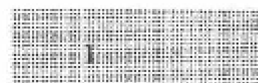
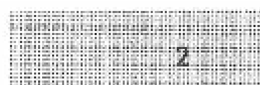
目录





第一章 绪 论

§1 数学地质学的发展简史

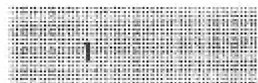
数学地质学是采用数学理论和方法，以计算机为主要技术手段，定量化、智能化、可视化地研究地质过程中所产生的地质现象和资源状况的一门地质学边缘学科。

1840 年至 1935 年，是数学在地质学中初步应用和在个别方面进行少量分散研究的时期。1840 年莱伊尔利用古生物化石的统计分析对第三系进行划分。1890 年皮尔逊(Karl Pearson)编写了《数学进化论贡献》丛书，内有古生物化石的统计分析。1914 年至 1934 年，列文生-列星格(Левинсон-Лессинг)通过考察岩石岩浆系数的频率分布，研究了安山岩、玄武岩、英安岩、流纹岩的分类。1929 年勃林克曼(R. Brinkmann)进行了一些生物地层学方面的统计研究工作。

1936 年至 1945 年，数学方法的应用范围由地质学个别问题逐渐扩展至地质学的一些分支。1939 年西姆波森(G. G. Simpson)等编著了《定量动物学》一书，为古生物统计学的发展奠定了基础。美国人克鲁拜因(W. C. Krumbein)从 1934 年开始进行沉积作用和地层的统计分析工作，成为美国数学地质学的奠基人。1944 年前苏联维斯捷列乌斯(A. Б. Вистелиус)在前苏联科学院报告集上发表了《分析地质学》一文，提出用定量方法研究地质学问题的初步思想。从此，他从事数学地质工作 30 余年，成为前苏联数学地质学的创始人和国际数学地质协会的第一任主席。

1946 年至 1960 年，数学方法应用于地质学的许多分支，单变量、双变量统计方法被普遍应用。前苏联已有人研究金属矿床元素的统计分布特点。1954 年绍(D. M. Shaw)等应用统计方法研究地球化学问题。1956 年初，切叶思(F. Chayes)应用均值、方差、标准差于岩石学研究中。1958 年克鲁拜因开始从事区域地层统计分析方面的工作。在此期间，电子计算机的应用和数字绘图仪的诞生，为地质学与数学的结合创造了条件。1958 年克鲁拜因首次在地质学杂志上公布电子计算机地质应用程序。

1961 年至 1970 年，数学方法和电子计算机在地质学中开始广泛应用。亚利桑那大学从 1961 年开始召开了一系列“电子计算机在矿产工业中的应用”讨论会。第二代计算机的成批生产和应用导致数学地质学文献数目激增。1964 年达特蒙斯大学第一次成功地应用了计算机分时系统，美国堪萨斯地质调查所召开了第一次“电子计算机在地球科学中的应用讨论会”，《美国石油地质工作者公报》杂志设立了“电子计算机应用”的专门编委。堪萨斯地质调查所从 1966 年开始连续出版电子计算机程序集。1967 年在美国石油地质工作者协会中建立了电子计算机数据存储和索取委员会。同年成立了国际地质科学联合会的地质数据储存、自动处理和索取委员会。1968 年在巴黎召开的国际地质会议上成立了国际数学地质协会，并开始出版《国际数学地质协会杂志》和《地质计算程序公报》；美国地质调查所首次公布其电子计算机贡献文集；电子计算机在地球科学应用方面的第一本书出版。在这一阶段，多元统计方法在地质学中大量应用，数学地质学发展成为一门独立的学科。



1971 年后，数学地质学科向更高水平发展。地质过程的数学模拟在数学地质学中占据愈来愈重要的地位，愈来愈多的数学方法应用于地质学中；地质统计学取得明显进展，由法语国家向英语国家逐渐推广，并且水平不断提高；地质多元统计有形成独立分支的趋势。数学和地质学的不断结合推动了数学地质学的进展。

20 世纪 70 年代末，中国成立了数学地质学分会，先后召开全国数学地质学术讨论会 6 次，专题学术会议 12 次，推动了我国数学地质学的发展。

§2　数学地质学的主要研究内容与方法

数学地质学主要研究概率论与数理统计、地质统计学、地质建模技术、模糊数学、灰色理论、非线性科学、图形处理技术等内容。

(1) 概率论与数理统计。主要包括随机样本分析、随机抽样技术、蒙特卡罗技术、趋势分析、回归分析、判别分析、因子分析、聚类分析、最优分析等，这部分内容以概率论和数理统计理论为基础，对地质数据进行统计分析，得到相应的统计结果并进行地质解释。这些常规的数学地质学方法在地质领域已有较广泛的应用。

(2) 地质统计学。是在地质分析和统计相互结合的基础上，用随机函数的形式体系来评价和探索区域化地质变量的理论和方法。主要包括变异理论、克立金技术等。

(3) 地质建模技术。是运用计算机技术，将空间信息管理、地质解释、空间分析和预测、地学统计、实体内容分析和图形可视化工具等结合起来，对地质对象以及相关的工程活动进行再现和分析的技术。主要包括模型与模拟、建模性质、模型分类、数字油田、勘探数据库与数据银行、地质模型等内容。

(4) 模糊数学。模糊性现象是一种普遍存在的现象，而模糊数学就是研究和处理这些模糊现象的一种数学理论和方法。其基本研究内容有三个方面，即模糊数学的理论以及它与精确数学、随机数学的关系。

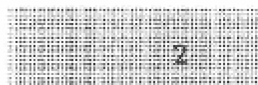
(5) 灰色理论。又称灰色系统理论，与研究“随机不确定性”的概率统计和研究“认知不确定性”的模糊数学相比，灰色理论的研究对象是具有“部分信息已知，部分信息未知”特点的小样本、贫信息的不确定系统。灰色系统主要研究方法包括灰色系统建模方法、灰色系统控制理论、灰色关联分析方法、灰色预测方法、灰色规划方法、灰色决策方法等。

(6) 非线性科学。非线性科学研究的是变量间存在的非线性关系。自然现象的复杂性决定了用非线性描述的必要性，它使所进行的描述更为合理可信。地质学中的非线性理论和方法包括分形理论、耗散结构理论、混沌理论、突变理论、协同学、非线性动力学等。

(7) 图形处理技术。是用计算机图形化手段对地质信息和研究成果进行展示和分析的一种技术，其主要内容包括图形算法、计算机显示技术等。随着计算机图形技术的发展，计算机绘图工作进展迅速，在 20 世纪 80 年代后期基本上发展成为一门独立的学科。

§3　未来展望

自然科学的一般发展趋势是由定性研究向定量研究发展，而数学地质学正是地质学由定性研究向定量研究发展的主要手段，虽然目前在地质学研究中并不占主导地位，但由于它和尖端科学技术在地质学中的应用密切相关，可以预见数学地质理论和方法必将在地质学研究中发挥十分重要的作用。今后数学地质学的发展主要有以下几个方向：



(1) 传统数学地质学方法的进一步改进与应用。以多元统计分析为主的传统数学地质学方法,目前已在许多地质定量研究中得以应用,事实也证明了这些方法的有效性。但这些传统数学地质学方法在取得成功的同时,也暴露出不少问题。如数学模型与地质概念模型的吻合程度因所研究的对象不同而有较大差异,甚至有些偏差过大,导致处理结果无法解释地质现象。虽然原因是多方面的,但数学模型本身的局限性也显而易见。今后,在合理完善数学模型的基础上,传统的数学地质学方法依然有强大的生命力。

(2) 非线性科学在地质学中的进一步应用。地质系统是一个复杂的巨系统,具有非平衡性、非线性、多尺度性、突变性、自组织性、自相似性、有序性和随机性等特点,因此,为了研究和解决地质系统的重大理论和实际问题,必须应用非线性理论和方法。一个被称为"非线性地质学"的新发展方向越来越引起人们的重视,将逐渐得到推广和应用。

(3) 地理信息系统(geographic information system,GIS)在地质学中的推广应用。地理信息系统是计算机软硬件支持下的空间数据输入、存储、检索、运算、显示和综合分析的应用技术系统;该系统的研究重点是空间实体及其相互关系,主要用途是分析和处理在一定地理区域中分布的各种现象和过程。近年来,地理信息系统已逐渐应用于地质研究。由于该系统自身的特点及优势,今后必将成为数学地质学研究的重要方向。

(4) 人工智能在地质学中的应用将进一步发展。除了继续建立各种地质专家系统外,将更重视多种人工智能技术的综合应用,特别是应用人工神经网络和遗传算法,使人工智能在地质学中的应用达到一个新的高度。

(5) 地质数据库共享技术。在计算机网络技术迅速发展的基础上,利用数据库共享技术,可以充分使用分布在各地的通用勘探数据库、图形库及数据银行中的数据资源,随时随地提供最新研究资料。

(6) 数学新理论新方法、计算机新技术的不断引进。对一些用目前的数学方法难以描述的地质现象,随着数学理论和计算机技术的发展,问题将得以合理解决,从而提高数学地质学方法的应用价值。

(7) "数字油田"技术的推广应用。利用测绘、数据库、地理信息系统、因特网、虚拟现实等技术,可将油田勘探、开发、生产过程中的资料采集、处理、存储、显示等环节实现数字化和可视化。结合分析和运算功能,可更加准确地预测油气资源的分布状况,实现决策的合理性及提高研究工作效率。

第二章　地质变量与地质数据

§1　地质变量

一、地质变量的概念及其分类

1. 地质变量的概念

地质变量是反映某地质现象在时间或空间上变化规律的量。如生油岩的厚度、地层的埋藏深度、生油岩中有机质的丰度等。

2. 地质变量的分类

造成地质现象的因素的复杂性，导致表示不同地质特征的地质变量各不相同。但是，可以根据它们所取数据的性质及方法，将其分为观测变量（定性变量和定量变量）和综合变量。

观测变量是可以直接进行观测、分析或度量的地质变量。如地层的厚度、石油的密度和粘度、岩石的颜色等。

综合变量是把两个或两个以上的观测变量按一定的方式进行组合而得到的具有综合意义的地质变量。如区分天然气成因类型的甲烷系数$M(M = C_1/\sum_{i=1}^{5} C_i)$，当$M > 99\%$时，认为是生物成因气，否则认为是热解成因气。又如有机质转化率（总烃与有机碳之比）。

3. 地质变量的观测值

用各种化学、物理以及直接观测的方法获得的地质变量的各种数据和其他形式记录的资料统称为地质变量的观测值。

二、地质变量的特征

1. 具有明确的地质意义

地质意义主要是指对地质变量所代表的特定研究对象的认识，主要包括：对地质变量所代表的石油地质特征的认识，如地层的时代、地层温度、圈闭闭合面积等；对地质变量所代表的盆地地球化学特征的认识，如有机质类型和丰度、干酪根成熟度等；对地质变量所代表的地球物理特征的认识等。

2. 具有明显的统计性质

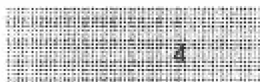
很多地质变量是随机变量，因此，它们的观测值具有明显的统计意义，如观测值的平均值是地质变量数学期望的估计值，而观测值的方差反映了地质变量在研究区域上的变异。

3. 具有相关性

地质变量之间具有一定程度的相关性，如岩石的渗透率与有效孔隙度密切相关。

三、地质变量的选择

分析研究地质变量的目的是想通过它们预测地质体的特征及有关的地下资源。用什么样的地质变量才能较好地实现研究目标，这就是地质变量的选择问题。例如，要想通过一些地质变量预测某沉积单元的油气资源量，就要选择与油气资源量相关的生油条件、储集条件、保存条件、圈闭条件等地质变量。一般来说，地质变量的选择应遵循以下基本原则：



(1) 基于地质概念模型。以相关地质学科理论为指导,分析、建立研究问题的地质概念模型,依据地质概念模型选择相应的地质变量。

(2) 结合地质变量间的相关性。地质变量之间存在着不同程度的相关性,应选择与矿藏形成或地质体特征等有密切联系的控矿因素和找矿标志。

(3) 地质变量要有代表性。地质变量的代表性是指所选择的地质变量能否较好地表征某地质作用过程的进行程度,或者变量的观测区与未观测区之间的相似程度。

(4) 地质变量要有明确的地质意义。拟定的地质变量,特别是构造的综合型地质变量要有确切的地质含义。如在油气资源评价中,生油岩体积与沉积岩体积之比表示评价区的生油条件,而近油源圈闭面积与沉积岩面积之比则表示评价区的圈闭条件,总烃与有机碳之比表示有机质转化率等。

§2 地质数据

一、地质数据的概念

用以代表地质体或其他自然产物特性的实物样子称为样品。地质样品的采集对象有岩体、地层、矿体、油气、生油岩、储集层、土壤及各种松散的沉积物、地表水及地下水、植物、空气等。用各种物理、化学方法以及直接观测的方法获得的用以表示样品特性的各种数据和其他形式记录的资料统称为地质数据或样品变量观测值。

二、地质数据的分类

地质数据是地质样品的变量观测值。因此,从狭义上讲地质数据分为定性数据和定量数据,从广义上讲它可以是定性数据、定量数据、图形或其他形式记录的资料等。根据地质数据的来源,地质数据分为观测、综合、经验数据三类。

1. 观测数据

指对样品(或采样对象)用各种物理、化学或直接观测的方法获得的表达样品(或采样对象)特性的数据。这种源于样品、没有经过任何加工处理的数据,又称为原始数据。依据数据的性质,又分为定性数据和定量数据两类。

(1) 定性数据。定性数据是指用符号或代码表示的没有数量概念的观测数据。可将其分为名义型和有序型两类:

① 名义型数据是没有数量概念和次序之分,但彼此之间有“相等”或“不相等”关系的定性数据。如岩石的红、绿、灰、黄色可以用字母 A,B,C,D 表示,又如砂岩、泥岩、灰岩可以用 S,N,H 代替,它们之间有 $A=A, A\neq B, S=S, S\neq N$ 的关系。

② 有序型数据是没有数量概念,但彼此之间具有次序关系的定性数据。如Ⅰ,Ⅱ,Ⅲ型干酪根可用数字1,2,3表示,它们之间有Ⅰ型干酪根的生烃潜力优于Ⅱ型干酪根的关系。

(2) 定量数据。定量数据是指用数值来描述的观测数据。可将其分为间隔型数据和比例型数据两类:

① 间隔型数据是有明确数量概念和地质含义的定量数据。如以基准海平面起算的地层分层数据就是典型的间隔型数据。它们之间具有相等、不等以及大于、小于关系,其差异具有实际的地质意义。如某地层底界和顶界分层深度值之差等于该地层的厚度。

② 定量数据的比值构成比例型数据。这类数据本身及它们的差值都有实际意义。比例型数据是大于等于0的实数组成的数据集合,这是它与间隔型数据的一个重要区别。如

两地层厚度的比值反映其中一个地层厚度是另一个地层厚度的百分之几，或者反映某种沉积环境，或者反映生油条件等。

2. 综合数据

综合数据是指由定量数据(或经定量化处理后的定性数据)经有限次算术运算后得到的定量数据。这种数据具有明显的地质意义，例如总烃含量、时间-温度指数、生油岩厚度与沉积岩厚度的比等。另外，随机变量的各种数字特征，如平均值、标准差、极差、相关系数等都可视为综合数据。

3. 经验数据

经验数据是在研究地质现象和规律的基础上，根据大量实际资料和经验总结归纳出的数据，如单储系数、排烃系数、聚集系数等。经验数据是大量地质信息的综合反映，地质意义明确，但它受哪些主控因素的影响，以及各因素之间的作用关系等问题目前尚不清楚。另外，经验数据还具有较明显的地域性。因此，在油气资源评价等工作中使用经验数据时，要特别注意对比地质条件的相似性。

三、地质数据的主要特点及数据矩阵

1. 地质数据的主要特点

由于地质系统、地质条件和地质作用的复杂性，测试手段的差异等，导致地质数据有如下几个主要特点：

(1) 地质数据类型多，性质不一，地质内涵丰富，量纲不统一，定量数据的数量级相差大，各类数据的数量和精度相差悬殊。

(2) 地质数据往往是多种地质因素综合作用的结果，故具有混合分布特征。

(3) 地质数据以定量数据为主，而定性数据的定量化研究和应用目前尚不成熟。

地质数据的特点决定了地质数据不是单一性质的数据集合，而是多种来源的混合数据集合，这一特点客观存在且不易改变。使用地质数据时，要注意它们的适用性，同时还要研究和改进数据加工和处理技术，发挥各种地质数据的作用，方可使地质研究获得良好的效果。

2. 数据矩阵

为便于数据处理，地质数据常用数据矩阵表示。假设有 n 个样品，每个样品有 m 个变量，那么样品变量的观测值可用以下数据矩阵 $\boldsymbol{X}$ 表示：

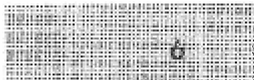
$$\boldsymbol{X}=(x_{ij})_{n\times m}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

式中　x_{ij}——第 i 个样品第 j 个变量的观测值。

常把 $\boldsymbol{X}$ 的第 j 列记为 X_j，它是第 j 个变量的 n 次观测值。有时也将数据矩阵记为：

$$\boldsymbol{X}=(x_{ij})_{m\times n}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

式中　x_{ij}——第 i 个变量的第 j 次观测值。



例如，地质圈闭的5个参数(表2-1)可以用5行4列的矩阵式(2-1)表示。

表2-1 地质圈闭数据

圈闭编号	闭合面积/(10^2 m^2)	闭合高度/m	长短轴比	埋藏深度/m
1	1 000	500	1.5	2 000
2	250	150	1.0	2 200
3	100	70	3.0	1 500
4	10	200	2.0	1 800
5	40	100	5.0	2 500

$$\boldsymbol{X} = (x_{ij})_{5\times 4} = \begin{bmatrix} 1\,000 & 500 & 1.5 & 2\,000 \\ 250 & 150 & 1.0 & 2\,200 \\ 100 & 70 & 3.0 & 1\,500 \\ 10 & 200 & 2.0 & 1\,800 \\ 40 & 100 & 5.0 & 2\,500 \end{bmatrix} \tag{2-1}$$

§3 地质数据的预处理

地质数据的预处理是指在定量研究地质问题时，预先对原始数据进行的各种处理。其主要内容为定量数据的标准化、定性数据的定量化、原始数据的网格化、原始数据的简缩和增补、离群数据的识别与剔除等。

一、定量数据的标准化

定量数据的标准化是对变量的观测值进行标准化。其目的是消除或抑制不同变量观测值数量级的巨大差异，使它们在同一尺度范围下参与地质研究。标准化方法有标准差标准化、极差标准化、极差正规化、总和标准化、最大值标准化、模标准化和中心标准化等。其中最常用的是标准差标准化、极差标准化和极差正规化。

1. 标准差标准化

标准差标准化是变量 X_j 的每个观测值 x_{ij} 减去观测值的平均值 $\bar{X}_j$，再除以观测值的标准差 S_j，即数据矩阵 $\boldsymbol{X}$ 中第 j 列上的每个元素减去该列元素的平均值，再除以第 j 列元素的标准差，最终得到变量 X'_j。变换公式为：

$$x'_{ij} = (x_{ij} - \bar{X}_j)/S_j \qquad (i = 1,2,\cdots,n; j = 1,2,\cdots,m) \tag{2-2}$$

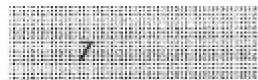
式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据，即第 i 个样品第 j 个变量的观测值)；

$\bar{X}_j$——第 j 个变量观测值的平均值，$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ (j = 1,2,\cdots,m)$；

S_j——第 j 个变量观测值的标准差，$S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{X}_j)^2}\ (j = 1,2,\cdots,m)$。

变量 X'_j 叫做标准化变量，其特点是平均值 $\bar{X}'_j = 0$，标准差 $S'_j = 1$，故变量 X'_j 又叫做规格化变量。



对式(2-1)中的数据标准差标准化后，得到新的数据矩阵：

$$X' = \begin{bmatrix} 1.949 & 1.916 & -0.707 & 0.000 \\ -0.081 & -0.350 & -1.016 & 0.587 \\ -0.487 & -0.867 & 0.354 & -1.468 \\ -0.731 & -0.026 & -0.354 & -0.587 \\ -0.650 & -0.673 & 1.768 & 1.468 \end{bmatrix}$$

2. 极差标准化

极差是第 j 个变量 X_j 观测值的最大值与观测值最小值的差，即：

$$\Delta X_j = \max_{1\leqslant i\leqslant n} x_{ij} - \min_{1\leqslant i\leqslant n} x_{ij} \qquad (j = 1,2,\cdots,m)$$

极差标准化是变量 X_j 的每一个观测值 $x_{ij}(i=1,2,\cdots,n)$减去 X_j 观测值的平均值 $\overline{X}_j$，再除以极差 ΔX_j。变换公式为：

$$x'_{ij} = (x_{ij} - \overline{X}_j)/\Delta X_j \qquad (i = 1,2,\cdots,n;j = 1,2,\cdots,m) \qquad (2\text{-}3)$$

式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据)；

$\overline{X}_j$——第 j 个变量观测值的平均值；

ΔX_j——第 j 个变量观测值的极差。

极差标准化后，变量 X'_j 的极差为 1。对式(2-1)中数据极差标准化后，得到新的数据矩阵：

$$X' = \begin{bmatrix} 0.727 & 0.688 & -0.250 & 0.000 \\ -0.030 & -0.126 & -0.375 & 0.200 \\ 0.182 & -0.312 & 0.125 & -0.500 \\ -0.273 & -0.009 & -0.125 & -0.200 \\ -0.242 & -0.242 & 0.625 & 0.500 \end{bmatrix}$$

3. 极差正规化

极差正规化的变换公式为：

$$x'_{ij} = (x_{ij} - \min_{1\leqslant i\leqslant n} x_{ij})/\Delta X_j \qquad (i = 1,2,\cdots,n;j = 1,2,\cdots,m) \qquad (2\text{-}4)$$

式中 x'_{ij}——标准化后的数据；

x_{ij}——标准化前的数据(原始数据)；

ΔX_j——第 j 个变量观测值的极差；

$\min\limits_{1\leqslant i\leqslant n} x_{ij}$——第 j 个变量观测值的最小值。

对式(2-1)中数据极差正规化后，得到新的数据矩阵：

$$X' = \begin{bmatrix} 1.000 & 1.000 & 0.125 & 0.500 \\ 0.242 & 0.186 & 0.000 & 0.700 \\ 0.091 & 0.000 & 0.500 & 0.000 \\ 0.000 & 0.302 & 0.250 & 0.300 \\ 0.030 & 0.070 & 1.000 & 1.000 \end{bmatrix}$$

由式(2-4)可知，变量 X'_j 最大值为 1，且 $x'_{ij}\geqslant 0$，即新数据在区间[0,1]内。

二、定性数据的定量化

定性数据的定量化是指把定性数据变换为数值。根据定性数据状态的多少，可分为二态和多态有序定性数据。两类定性数据的定量化方法都是对定性数据的状态赋值。

1. 二态定性数据的变换

只有两种对立状态的定性数据为二态定性数据。可用 0 和 1 表示这两个状态，从而实现定性数据的定量化。如某观测点有无某种化石，就只有两种可能，若有则用 1 表示，若无就用 0 代表。一般来说，按以下原则处理：

二态定性数据	状 态	肯定或有利	否定或不利
	赋 值	1	0

2. 多态有序定性数据的变换

多态有序定性数据是指状态多于两个，并且状态又可按一定次序排列的定性数据。如储层岩心的含油性，按含油程度可分为四级，采用等差方式赋值如下：

四态有序	状 态	不含油	油 斑	含 油	饱含油
定性数据	赋 值	0	1	2	3

又如，按颜色可将泥岩分为四级，为区分各级泥岩的生油能力，可采用非等差方式赋值如下：

四态有序	状 态	红 色	浅灰色	灰 色	黑 色
定性数据	赋 值	0	1	3	5

一般按以下原则处理：

多态有序	状 态	状态 1	状态 2	状态 3	…
定性数据	赋 值	x_1	x_2	x_3	…

三、原始数据的网格化

原始数据的网格化是指把平面上无规则分布的定量数据 z_i 分配到矩形网格的每个交点上（图2-1），产生规则分布的定量数据。网格化的方法很多，如全点插值法、圆内插值法、曲面插值法、克立金法等。在此仅介绍既简单又实用的按象限取点距离加权平均法。

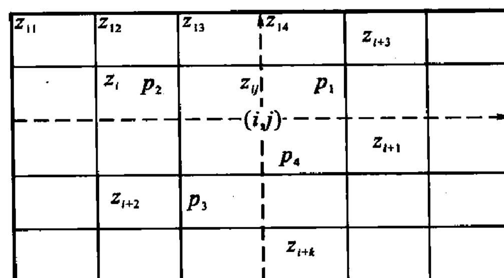


图 2-1 原始数据网格化示意图

以网格交点(i,j)为原点建立坐标系（图 2-1），在各象限内各取一个距点(i,j)最近的数据点，记为 $p_i(i=1,2,3,4)$，各点上相应的数据值分别为 $z_i(i=1,2,3,4)$，可以用 p_i 点上的数据值 z_i 的算术平均值

$$\bar{z}=\frac{1}{4}\sum_{i=1}^{4}z_i$$

作为网格交点(i,j)上的估计值z_{ij}。

考虑到数据点距网格交点(i,j)越近，对网格点的估计值影响越大，因此取距离的倒数作为权求网格交点(i,j)的估计值。

假设数据点$p_i(i=1,2,3,4)$到网格交点(i,j)的距离为$d_i(i=1,2,3,4)$，那么网格交点(i,j)上的估计值为：

$$z_{ij}=\sum_{i=1}^{4}\frac{z_i}{d_i}\Big/\sum_{i=1}^{4}\frac{1}{d_i} \tag{2-5}$$

在按式(2-5)计算z_{ij}的过程中，当出现$d_i=0$时，则以z_i作为网格点(i,j)上的估计值。

对于某些网格点(如边界网格点)，不能在四个象限中都找到数据点，则在有数据点的象限内取距离近的点上的数据进行加权平均(至少有一个数据)。在按象限取点距离加权平均插值法中，每个象限内也可以取多个距插值点近的数据点进行加权平均，其过程与上类似。

对每个网格交点进行上述计算，即可完成对原始数据的网格化工作。

【例 1】如图 2-2 所示，已知四个数据点$p_1(4,4)$，$p_2(1,4)$，$p_3(1,2)$和$p_4(5,2)$，各点上的数据值依次为 3,2,2,2，求插值点$p(3,3)$上的估计值。

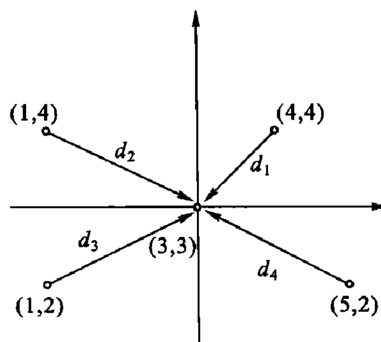


图 2-2　点的插值

解：

① 各点到点$p(3,3)$的距离：$d_1=\sqrt{2}$，$d_2=d_3=d_4=\sqrt{5}$。

② 式(2-5)的分母：$\sum_{i=1}^{4}\frac{1}{d_i}=2.0487$。

③ 式(2-5)的分子：$\sum_{i=1}^{4}\frac{z_i}{d_i}=4.8045$。

④ 插值点的值：4.804 5÷2.048 7 =2.345 1。

四、原始数据的简缩和增补

1. 原始数据的简缩

当分布在研究区上的地质数据点很多(可能出现反映相同地质特征的多个近似数据点)时，或者是数据在研究区上的分布极不均匀时，不仅会使计算量增加，而且也无助于最终的成果解释，甚至在计算过程中还会出现不可预料的计算病态问题。因此，就需要对作用不大或相近、可有可无的多余数据予以舍弃，这就是数据的简缩。

数据的简缩方法一般包括分区加权平均法、分区滑动平均法和随机删点法。

(1) 分区加权平均法。

假设研究区内每个地质数据点有m个变量，根据实际需要将研究区划分成大小相等或不等的n个小区，并且每个小区内至少有一个数据点，那么第j个小区内第i个数据点上第k个地质变量的观测值为$z_{jki}(j=1,2,\cdots,n;k=1,2,\cdots,m;i=1,2,\cdots,n_j)$，而第$j$个小区内第$k$个地质变量的简缩值为：

$$z_{jk}=\frac{1}{n_j}\sum_{i=1}^{n_j}z_{jki}\qquad(j=1,2,\cdots,n;k=1,2,\cdots,m) \tag{2-6}$$

式中　z_{jk}——第j个小区第k个变量观测值的简缩值；

n_j——第j个小区地质数据点数；

z_{jki}——第j个小区第i个数据点上第k个变量的观测值。

按照式(2-6)对研究区内原始数据进行处理后,相当于每个小区内有一个有效数据点,从而将原来大量的数据点简化为 n 个有效数据点。

(2) 分区滑动平均法。

分区滑动平均法的分区方法和分区原则与分区加权平均法相同,但这种方法要考虑简缩后数据点的位置。

如果第 j 个小区内有 n_j 个数据点,每个数据点上有 m 个地质变量的观测值,其中第 i 个数据点的坐标为(x_{jki}, y_{jki}),那么第 j 个小区简缩后的有效数据点坐标值及变量由式(2-7),(2-8)给出:

$$\begin{cases} x_{jk} = \sum_{i=1}^{n_j} x_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \\ y_{jk} = \sum_{i=1}^{n_j} y_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \end{cases} \tag{2-7}$$

$$z_{jk} = \sum_{i=1}^{n_j} z_{jki} / n_j \tag{2-8}$$

$$(j = 1,2,\cdots,n; k = 1,2,\cdots,m)$$

式中 x_{jk}, y_{jk}——第 j 个小区第 k 个地质变量观测值简缩后的横坐标和纵坐标;

z_{jk}——第 j 个小区第 k 个地质变量的简缩值;

x_{jki}, y_{jki}——第 j 个小区第 k 个地质变量观测值的第 i 个数据点的横坐标与纵坐标;

z_{jki}——第 j 个小区第 k 个地质变量观测值的第 i 个数据;

n_j——第 j 个小区地质数据点数。

按上述公式算出的坐标有 m 个,如果需要一个统一的坐标点,则可根据地质变量观测值的大小,采用加权平均的方法算出。另外,根据实际需要,也可采用其他的计算方法。

(3) 随机删点法。

对于探区内的局部数据点密集区,随机删去一些数据点,既可减少计算工作量,又可提高计算过程的稳定性。删除点的方法是对数据点编号,用随机抽样法删去其中的一些数据点。

2. 数据的增补

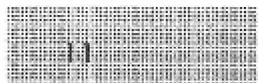
在一般情况下,探区内投入的工作量是不均匀的,特别是勘探早期阶段。因此在区域上会出现数据点空白区,在这种空白区往往需要补充一些数据点,这就是数据的增补。

在数据点空白区补充数据点时,可以用临近数据点上的数据外推,即根据数据的变化趋势补充适量的数据点,也可以用某种插值方法补充一定数量的数据点。值得注意的是:补点的目的是为了全区计算的稳定性,而原空白区的计算结果仅供参考。

此外,对于多变量的地质样品,由于分析化验项目不一定完全一致或其他原因,导致某些样品缺少某些变量的观测值。对于那些研究需要而又缺少观测值的变量,可以用该变量邻近区域上观测值的平均值作为该变量观测值的近似值。

五、离群数据的识别与剔除

相对研究区的观测数据来说,局部的异常高值和异常低值称为离群数据。这种数据往往直接影响到基于观测数据的数据处理过程和对计算结果的合理解释。对于某些已知因素造成的离群数据,可进行相应的数据校正。如在油气地表化探中,由地表自然地理条件、土



壤的类型和颜色等导致的有关指标含量的差异，经过相应的校正，即可以消除数据中的干扰。

如果离群数据是地质现象的真实反映，当它们对数据处理过程和计算结果产生消极影响时，应对这些数据进行适当处理。对于那些人为等因素造成的错误数据，理所应当地删除或重新进行观测。然而，判断造成离群数据的因素是很困难的。在实际工作中，总是假设数据是真实的，在此假设下讨论对离群数据的挑选和处理。

对离群数据进行处理的第一步工作是挑选离群数据，这就涉及离群数据的界限问题。下面简单介绍离群数据的界限确定和处理方法。

1. 类比法

以实际工作经验确定离群数据的界限，以此界限识别区域上的离群数据。斯米尔诺夫根据实际经验，总结出确定矿床品位离群数据的界限（表 2-2，其中离群品位高出平均品位的倍数项是经验数据）。

从矿床成因角度看，绝大多数的油气藏都属于与沉积岩有关的矿床。所以，确定油气勘探、开发中石油地质数据界限时，可以参照表 2-2 中的Ⅰ，Ⅱ矿床类型。

表 2-2　矿床品位离群数据的界限

矿床类型	组分分布性质	典型矿床	离群品位高出平均品位的倍数
Ⅰ	很均匀	大多数沉积矿床	2～3
Ⅱ	均　匀	复杂沉积矿床与变质矿床	4～5
Ⅲ	不均匀	绝大多数有色金属矿床	8～10
Ⅳ	很不均匀	大多数稀有金属矿床和金矿床	12～15
Ⅴ	极不均匀	某些稀有金属矿床和金矿床	＞15

2. 计算法

用经验公式确定离群数据的界限。沃洛多莫夫给出的计算离群数据界限的经验公式为：

$$ch = c_1 + (n-1)c_1(c_1 - c_2)/c_2 \tag{2-9}$$

式中　ch——正常数据的最大值，大于 ch 的数据即为离群数据；

c_1——校正前（包括离群数据）的样品平均值；

c_2——校正后（不包括离群数据）的样品平均值；

n——样品总数（包括离群数据）。

在一组数据中，离群数据一般只有少数几个，当个数太多时就不应是离群数据。在实际计算时，令$(c_1-c_2)/c_2=20\%\sim30\%$，由式(2-9)可求出离群数据的界限值。此界限值显然与样品数 n 有关。n 越大，ch 的偏离可能越大，故此法适合对小子样的检验。

3. 统计检验法

在观测数据来自同一个总体的前提下，统计检验法的思路是检验数据是否服从正态分布。若通过检验，则认为数据中不存在离群数据，否则认为数据中存在离群数据，这时就需要识别出其中的离群数据并对其进行有关的处理。统计检验的关键是构造一个合适的统计量及其所服从的分布，在此基础上确定相应的假设检验方法。

(1) 正态分布的 χ^2 检验法。

对于来自正态总体的 n 个观测数据 $x_i(i=1,2,\cdots,n)$，将区间$(-\infty,+\infty)$划分为 m 个区间：

$$(a_0,a_1),(a_1,a_2),\cdots,(a_{m-1},a_m) \qquad a_0=-\infty,a_m=+\infty$$

设 v_i 为数据落入第 i 个区间内的个数(频数)，p_i 为数据落入第 i 个区间内的理论概率。

假设 H_0：观测数据来自正态总体。

如果 H_0 为真，由皮尔逊定理知，统计量

$$\eta=\sum_{i=1}^{m}(v_i-mp_i)/(mp_i) \qquad (m>3) \tag{2-10}$$

服从 χ^2_{m-3}分布。

给定检验水平 α，查 χ^2 分布表得接受或拒绝 H_0 的临界值 $\chi^2_{m-3}(\alpha)$，若 $\eta<\chi^2_{m-3}(\alpha)$，则接受假设 H_0，否则拒绝假设 H_0，存在离群数据。

在对地质观测数据进行正态检验时，可将区间$(\min\limits_{1\leqslant i\leqslant n}x_i,\max\limits_{1\leqslant i\leqslant n}x_i)$均匀划分为 m 个区间：

$$(a_0,a_1),(a_1,a_2),\cdots,(a_{m-1},a_m)$$

其中 $a_0=-\infty,a_1=\min\limits_{1\leqslant i\leqslant n}x_i,a_{m-1}=\max\limits_{1\leqslant i\leqslant n}x_i,a_m=+\infty$。区间个数 m 可随数据点的增加而适当增加，一般取 $10\leqslant m\leqslant 40,\alpha=0.10,0.05,0.01$ 等。

(2) 正态分布的偏度和峰度检验法。

随机变量 X 的偏度 E_p、峰度 E_f 是指 X 的标准化变量$(X-\mu)/\sigma$ 的三阶中心矩和四阶中心矩：

$$E_p=E[(\chi-\mu)/\sigma]^3, \qquad E_f=E[(X-\mu)/\sigma]^4$$

对于观测数据 $x_i(i=1,2,\cdots,n)$，可计算出该批数据的偏度 E_p、峰度 E_f 的矩估计：

$$U_p=U_3/S^3, \qquad U_f=U_4/S^4$$

式中 $U_3=\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^3$ (样本三阶中心矩)；

$U_4=\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^4$ (样本四阶中心矩)；

$S=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$ (样本标准差)；

$\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$ (样本数据平均值)。

假设 H_0：观测数据来自正态总体。

若 H_0 为真，由概率统计理论可证，当 n 充分大时近似有：

$$U_p\sim N(0,S_p), \qquad U_f\sim N(3-6/(n+1),S_f)$$

$$S_p=6(n-2)/[(n+1)(n+3)]$$

$$S_f=24n(n-2)(n-3)/[(n+1)^2(n+3)(n+5)]$$

即

$$U_p/\sqrt{S_p}\sim N(0,1) \qquad [U_f-3+6/(n+1)]/\sqrt{S_f}\sim N(0,1)$$

因此，确定正态分布偏度、峰度检验方法如下：

① 对观测数据 $x_i(i=1,2,\cdots,n)$，求出其偏度和峰度的矩估计 U_p,U_f。

② 对给定的检验水平 α，求出检验临界值：

$$P_p = Z_{\alpha/2} \cdot \sqrt{S_p}, \qquad P_f = Z_{\alpha/2} \cdot \sqrt{S_p} + 3 - 6/(n+1)$$

若$|U_p| < P_p$且$|U_f| < P_f$，则接受假设H_0，认为该批数据服从正态分布，否则拒绝假设H_0，认为该批数据不服从正态分布，其中存在离群数据。

(3) 统计检验法中离群数据的界限。

对随机变量X，若其均值为μ，方差为σ^2，则由契比雪夫不等式知：

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geqslant 0.888\,9, \qquad P(\mu - 3\sigma < X < \mu + 3\sigma) \geqslant 0.937\,5$$

若随机变量$X \sim N(\mu, \sigma^2)$，则有：

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geqslant 0.954\,4, \qquad P(\mu - 3\sigma < X < \mu + 3\sigma) \geqslant 0.997\,4$$

因此，对于观测数据$x_i (i=1,2,\cdots,n)$，可确定它以大概率落入区间：

$$(\bar{x} - 2S, \bar{x} + 2S) \text{ 或 } (\bar{x} - 3S, \bar{x} + 3S)$$

若某个数据不在上述区间内，则认为它是离群数据，否则认为它是正常数据。

4. 对离群数据的处理

无论用哪种方法识别出的离群数据，都要对其进行处理。常用的处理方法有三种：

(1) 用离群数据邻近的正常数据插值或用某种平均值代替离群数据；

(2) 将离群数据缩放为上述区间的边界值，即对数据离群程度进行抑制；

(3) 剔除离群数据。

究竟采用哪种处理方法，要结合数据的实际情况而定。

5. 方法实施步骤

以统计检验法为例，讨论对离群数据的处理步骤。由于某些地质观测数据的离散程度较高，按前述方法对数据进行一次处理后，新的数据可能仍不能满足正态分布的要求，这时必须在新数据的基础上再进行同样的处理，反复进行多次，直至数据满足正态分布为止。因此，可将对离群数据的识别和处理归纳为一个迭代过程。对正态总体下的观测数据$x_i (i=1,2,\cdots,n)$处理步骤如下：

(1) 输入全部观测数据$x_i (i=1,2,\cdots,n)$，给定检验水平α。

(2) 在检验水平α下，对数据进行正态分布检验（χ^2检验或偏度、峰度检验），若通过检验则结束，否则进行下一步。

(3) 识别离群数据并对其进行相应处理，形成新数据。

(4) 重复(2)～(3)，直至数据通过正态分布检验。

计算机处理离群数据的流程如图 2-3。

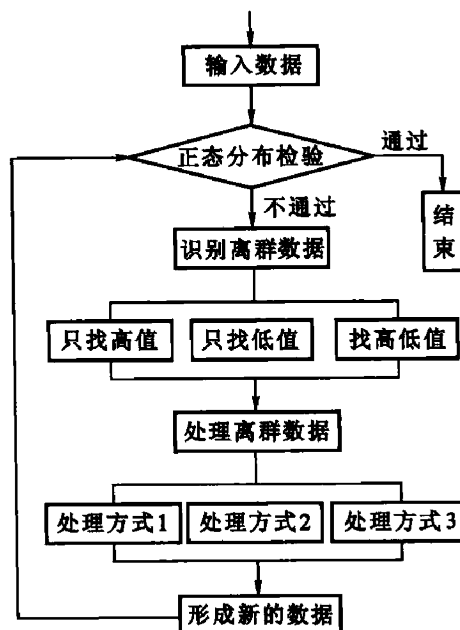


图 2-3　离群数据处理流程图

思考与练习

1. 什么是地质变量？地质变量主要有哪几种？地质变量有什么特征？
2. 简述地质数据的概念及其分类。
3. 地质数据有什么特点？
4. 简述地质数据矩阵的一般形式。
5. 什么是地质数据的预处理？为什么要对地质数据进行预处理？
6. 简述对地质数据进行标准化的常用方法、变换公式及变换后的数据特点。
7. 怎样把定性数据转化为定量数据？
8. 试述对原始数据进行网格化、简缩和增补的目的和方法。
9. 何谓离群数据？如何识别和处理离群数据？

10. 三个地质圈闭的有关参数观测值构成混合型数据(表 2-3)，试把此混合型数据写成矩阵形式并处理为[0,1]区间上的定量数据。

表 2-3　地质圈闭参数数据

圈闭编号	闭合面积/(10^2 m^2)	闭合高度/m	埋藏深度/m	有无断层	有无火成岩侵入
1	200	100	2 000	无	无
2	150	50	1 700	有	无
3	180	20	1 500	有	有

第三章 回归分析

§1 回归分析及其解决的问题

一、回归分析

在地质学研究领域，有很多地质变量之间的关系是相互依赖和相互制约，但又不能用一个方程将它们之间的变化关系表达出来。变量间的上述关系称为相关关系，存在这种关系的变量称为相关变量，如一个含油气盆地中的油气资源量 Q 随着盆地内生油岩的体积 V_1、储集岩的体积 V_2、近油源圈闭面积 S 的增大以及有机质转化率 k 的升高而增多，而随着盆地所经受的剥蚀次数 n 的增多而减少，即 Q 与 V_1,V_2,S,k,n 是相关变量。

一般说来，若变量 y 与 $x_i(i=1,2,\cdots,m)$ 是相关变量，又有它们的 n 组观测值，那么回归分析就是根据相关变量的 n 组观测值，研究它们之间的相关形式，确定它们之间的近似定量关系的一种多元统计分析方法。

二、回归分析解决的问题

由回归分析的概念可知，回归分析的研究对象是相关变量。因此，在回归分析中应解决变量间是否存在相关性，各变量间的相关程度、相关形式等问题，最终建立相关变量间的近似定量关系，即统计意义上的方程。

§2 多元线性回归分析

一、回归模型与回归方程

若变量 y 与 $x_i(i=1,2,\cdots,m)$ 之间具有

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \varepsilon \tag{3-1}$$

的关系，则称 y 与 x_i 之间具有 m 元线性相关关系，简称为线性关系，并称式(3-1)为线性回归模型，其中 $a_0,a_1,\cdots,a_m$ 为待定系数，ε 是误差项，并且 $\varepsilon \sim N(0,\sigma^2)$。

假设 $b_0,b_1,\cdots,b_m$ 是 $a_0,a_1,\cdots,a_m$ 的最佳估计值，那么式(3-1)可以改写为：

$$\hat{y} = b_0 + \sum_{i=1}^{m} b_i x_i \tag{3-2}$$

式(3-2)叫做 x_i 对 y 的线性回归方程，其中 $b_0,b_1,\cdots,b_m$ 叫做回归系数，$\hat{y}$ 叫做 y 的回归值。

二、确定回归系数

回归分析的任务之一就是确定回归系数。假设已有 y 和 $x_i(i=1,2,\cdots,m)$ 的 n 组观测值

$$(x_{1k},x_{2k},\cdots,x_{mk},y_k) \qquad (k=1,2,\cdots,n) \tag{3-3}$$

把式(3-3)中的 x_{ik} 代入式(3-2)，可得：

$$\hat{y}_k = b_0 + \sum_{i=1}^{m} b_i x_{ik} \tag{3-4}$$

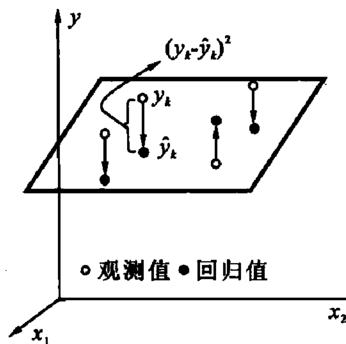


图 3-1 偏差示意图

确定回归系数的原则是使观测值 y_k 与回归值 $\hat{y}_k$ 的偏差平方和

$$Q_1 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \tag{3-5}$$

达到最小(图 3-1)。其中 Q_1 是以 $b_0, b_1, \cdots, b_m$ 为未知数的二次函数,并且 $Q_1 > 0$,根据极值原理有:

$$\frac{\partial Q_1}{\partial b_j} = 0 \quad (j = 0,1,\cdots,m) \tag{3-6}$$

把式(3-6)简化整理得:

$$b_0 = \bar{y} - \sum_{i=1}^{m} b_i \bar{x}_i \tag{3-7}$$

$$\sum_{j=1}^{m} s_{ij} b_j = s_{iy} \quad (i = 1,2,\cdots,m) \tag{3-8}$$

式中

$$s_{ij} = \sum_{k=1}^{n} x_{ik} x_{jk} - n\bar{x}_i \bar{x}_j, \qquad s_{iy} = \sum_{k=1}^{n} x_{ik} y_k - n\bar{x}_i \bar{y}$$

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik}, \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

由式(3-8)解出 $b_1, \cdots, b_m$,再由式(3-7)解出 b_0,即可得到回归方程式(3-2)。

三、回归检验

回归检验解决的是 y 与 x_i 之间是否具有线性关系的问题。为此,先定义以下几个统计量:

总离差平方和 $Q = \sum_{k=1}^{n} (y_k - \bar{y})^2$,

偏差平方和 $Q_1 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$,

回归平方和 $Q_2 = \sum_{k=1}^{n} (\hat{y}_k - \bar{y})^2$。

可以证明:

$$Q = Q_1 + Q_2 \tag{3-9}$$

Q, Q_1, Q_2 的自由度 f_Q, f_{Q_1}, f_{Q_2} 分别为 $f_Q = n-1, f_{Q_1} = n-m-1, f_{Q_2} = m$,并且满足等式 $f_Q = f_{Q_1} + f_{Q_2}$。

1. 复相关系数检验

由式(3-9)可知,Q_1 越小,Q_2 就越接近于 Q,说明变量 y 与 $x_i (i=1,2,\cdots,m)$ 的线性关系越密切,回归模型式(3-1)的偏差就越小,即回归方程式(3-2)所代表的变化关系就越接近实际。根据上述分析,可取 Q_2/Q 作为衡量回归方程显著性的一个指标。定义

$$r = (Q_2/Q)^{1/2}$$

为变量 y 与 $x_i (i=1,2,\cdots,m)$ 的复相关系数。r 的绝对值越接近于 1,变量间的相关性就越

密切，求得的回归方程就越显著。

2. *F* 分布检验

假设 H_0：变量 y 与 $x_i(i=1,2,\cdots,m)$ 之间没有式(3-1)给出的线性关系。

若 H_0 为真，那么 Q_1 就比较大，Q_2 就比较小。当 Q_2/Q_1 小于某个临界值时，就接受原假设 H_0；否则，就否定原假设 H_0，即变量 y 与 $x_i(i=1,2,\cdots,m)$ 之间有密切的线性关系。可以证明统计量

$$F=(Q_2/f_{Q_2})/(Q_1/f_{Q_1})=(Q_2/m)/[Q_1/(n-m-1)] \qquad (3\text{-}10)$$

服从第一自由度为 m、第二自由度为 $(n-m-1)$ 的 F 分布。对于给定的检验水平 α，在 $F_\alpha(m,n-m-1)$ 分布表上查得临界值 F_α。当 $F>F_\alpha$ 时，否定原假设 H_0，这时称回归方程是显著的，可以付诸应用；否则，接受原假设 H_0，即求得的回归方程不能应用。

四、非线性回归分析

变量间的相关关系并非都是线性的，如岩石渗透率 k 与声波时差 Δt、自然伽玛相对值 ΔGR 之间有

$$\ln k=a_0+a_1\ln\Delta t+a_2\Delta GR$$

的关系。对于这种情况，先用变量替换的方法将其转化为线性关系，然后再求线性回归方程。若令

$$y=\ln k,\qquad x_1=\ln\Delta t,\qquad x_2=\Delta GR$$

则上述问题就转化为

$$y=a_0+a_1x_1+a_2x_2$$

的线性回归分析了。

五、回归预测与控制

1. 回归预测

所谓回归预测就是把变量的给定值 $x_{ir}(i=1,2,\cdots,m;r=1,2,\cdots,n)$ 代入式(3-2)求得 y_r 的估计值：

$$\hat{y}_r=b_0+\sum_{i=1}^{m}b_ix_{ir}$$

当 $\min\limits_{1\leqslant k\leqslant n}x_{ik}\leqslant x_{ir}\leqslant\max\limits_{1\leqslant k\leqslant n}x_{ik}$ 时，y_r 落在区间 $(\hat{y}_r-\hat{\sigma},\hat{y}_r+\hat{\sigma})$，$(\hat{y}_r-2\hat{\sigma},\hat{y}_r+2\hat{\sigma})$ 内的概率分别为 0.68 和 0.95(图 3-2)。其中

$$\hat{\sigma}=\sqrt{Q_1/(n-m-1)}$$

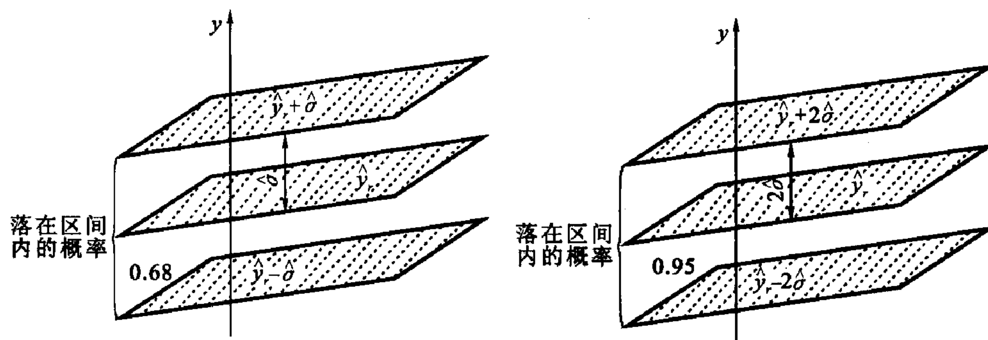


图 3-2　y_r 取值范围示意图

2. 控制

所谓控制就是调整 x_i 的值，使 y 落在期望区间 (y_1, y_2) 内。如改变影响原油采收率的储层非均质性、储层表面润湿性、流度比（驱动液流度与被驱动液流度的比值）等因素，使原油采收率提高至某个范围内，就是控制的典型例子。

§3 逐步回归分析

一、逐步回归分析的提出及其基本思想

1. 逐步回归分析的提出

(1) 变量 x_i 与 y 的相关程度不同。

众所周知，生油层的温度、埋藏深度制约了有机质热演化生油所需要的演化时间，但它们对演化时间的制约程度却不同。根据表 3-1 中的数据，求得生油门限时间 t 与生油层温度 T 和埋藏深度 H 的相关系数分别为 −0.662 7 和 −0.358 7。相关系数表明：有机质向石油演化所需要的时间随着地温的升高和埋藏深度的加大而缩短，其中温度对演化时间起着主导作用。

表 3-1　18 个盆地（地区）的生油层数据

序 号	含油气盆地（地区）	T/℃	H/m	t/Ma
1	杜阿拉盆地（喀麦隆）	65	1 200	70
2	落山矶盆地（美国）	115	2 400	12
3	文吐拉盆地（美国）	127	2 740	12
4	巴黎盆地（法国）	60	1 400	180
5	阿启坦盆地(1)（法国）	90	3 300	112
6	阿启坦盆地(2)（法国）	72	2 500	135
7	卡马尔圭盆地（法国）	106	3 250	38
8	阿尤恩地区	85	2 740	105
9	苏绿海盆地（沙巴）	120	3 050	12
10	塔拉纳基盆地（新西兰海上）	80	2 900	70
11	亚马逊盆地（委内瑞拉）	62	1 750	359
12	塔拉纳基盆地（新西兰海上）	95	3 350	32
13	东营盆地	93	2 200	35
14	潜江盆地	90	2 200	35
15	松辽盆地(1)	70	1 330	110
16	松辽盆地(2)	65	1 230	100
17	松辽盆地(3)	63	1 180	90
18	辽河盆地	81	1 700	50

上述实例表明：对拟定的变量 $x_i(i=1,2,\cdots,m)$ 来说，它们与 y 的相关程度不同，对 y 的作用也就不同，其中很可能有对 y 不起作用的变量，由此提出了"筛选"对 y 作用大的变量建立回归方程的逐步回归分析。

(2) 变量间的相关性。

地质现象是地质作用过程叠加的结果。因此描述地质现象的变量 $x_i(i=1,2,\cdots,m)$ 就

既有相对的独立性，又存在着一定的成因联系。对具有成因联系的一些变量，貌似各自对 y 都有不可忽视的影响，但当把其中一个变量选入回归方程后，又使得先选入的另一个变量对 y 的作用变得无足轻重了。因此，也要对已选入回归方程中的变量进行逐步"筛选"，这是提出逐步回归分析的另一个原因。

此外，还要考虑回归方程的简便和实用。

2. *逐步回归分析的基本思想*

逐步回归分析的基本思想是：在回归分析过程中，按变量 $x_i(i=1,2,\cdots,m)$ 对 y 作用的大小，把作用达到一定程度的变量 $x_r(1\leqslant r\leqslant m)$ 逐个"引入"回归方程，同时逐个检验已引入回归方程的变量 $x_a(x_a\in x_r)$ 对 y 的作用，如果它对 y 的作用已不显著，则再从回归方程中"剔除"x_a。如此进行下去，直到既没有对 y 作用显著的变量可引入回归方程，又没有作用不显著的变量从回归方程中"剔除"为止。

回归结束时，若引入的变量为 $x_{k_1},x_{k_2},\cdots,x_{k_l}$，则回归方程为：

$$\hat{y}=b_0+b_{k_1}x_{k_1}+b_{k_2}x_{k_2}+\cdots+b_{k_l}x_{k_l} \tag{3-11}$$

式(3-11)中就仅包含了对 y 作用显著的变量。

二、变量 x_i 对 y 的作用及作用大小的检验

在逐步回归分析中，要不断地判断变量 x_i 对 y 的作用，那么如何衡量变量 x_i 对 y 作用的大小呢？为此，需要构造一个衡量变量 x_i 对 y 作用大小的指标以及检验作用大小的方法。

1. *衡量变量 x_{k_a} 对 y 作用的指标*

对于引入了 l 和 $l+1$ 个变量的回归方程来说，参照式(3-9)有：

$$Q=Q_1^{(l)}+Q_2^{(l)},\qquad Q=Q_1^{(l+1)}+Q_2^{(l+1)}$$

由上式可得：

$$\Delta Q=Q_2^{(l+1)}-Q_2^{(l)}=Q_1^{(l)}-Q_1^{(l+1)} \tag{3-12}$$

式(3-12)表明，回归方程中增加一个变量后，导致回归平方和、偏差平方和的变化，并且回归平方和的增加量等于偏差平方和的减少量。通常称 ΔQ 为第 $l+1$ 个变量对 y 的方差贡献。一般情况下，变量 x_{k_a} 的方差贡献记为 V_{k_a}，它是衡量变量 x_{k_a} 对 y 作用大小的一个指标。

2. *检验变量 x_{k_a} 对 y 作用大小的方法*

(1) 检验 x_{k_a} 是否引入回归方程。

假设 H_0：变量 x_{k_a} 对 y 作用不显著，统计量

$$F_{k_a}=[V_{k_a}(x_{k_1},x_{k_2},\cdots,x_{k_l})/1]/[Q_1(x_{k_1},x_{k_2},\cdots,x_{k_l},x_{k_a})/(n-l-2)] \tag{3-13}$$

服从 $F(1,n-l-2)$ 分布。其中 n 为样本容量(数据组数)，l 为回归方程中的自变量个数。

给定检验水平 α，查 $F_\alpha(1,n-l-2)$ 分布表，得一个临界值 F_1。若 $F_{k_a}>F_1$，则否定假设 H_0，即变量 x_{k_a} 对 y 作用显著，应引入回归方程；否则，接受假设 H_0，即不引入 x_{k_a}。

(2) 检验 x_{k_a} 是否从回归方程中剔除。

假设 H_0：变量 x_{k_a} 对 y 作用不显著，统计量

$$F'_{k_a}=[V_{k_a}(x_{k_1},x_{k_2},\cdots,x_{k_l})/1]/[Q_1(x_{k_1},x_{k_2},\cdots,x_{k_l})/(n-l-1)] \tag{3-14}$$

服从 $F(1,n-l-1)$ 分布。

给定检验水平 α，查 $F_\alpha(1,n-l-1)$ 分布表，得一个临界值 F_2。若 $F'_{k_\alpha}>F_2$，则否定假设 H_0，即变量 x_{k_α} 对 y 作用显著，应留在回归方程中；否则，接受假设 H_0，即从回归方程中剔除变量 x_{k_α}。

三、实现逐步回归分析的变换公式

逐步回归是在多元回归的基础上派生出的计算技巧，它是通过对变量的相关系数增广矩阵实施一系列矩阵变换来实现逐步引入和剔除变量，求解回归方程。

1. *相关系数增广矩阵*

对变量进行标准差标准化，并将处理后的新变量仍然记为 $x_i(i=1,2,\cdots,m+1)$，其中的 $x_{m+1}=y$。对于标准化变量，可以证明回归系数 b'_j 满足方程组：

$$\sum_{j=1}^{m} r_{ij}b'_j = r_{im+1} \quad (i=1,2,\cdots,m) \tag{3-15}$$

式中 r_{ij}——原始变量 x_i 与 x_j 的相关系数。

把式(3-15)的系数矩阵增加一行一列，得相关系数增广矩阵：

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1m+1} \\ r_{21} & r_{22} & \cdots & r_{2m} & r_{2m+1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{m,m+1} \\ r_{m+1,1} & r_{m+1,2} & \cdots & r_{m+1,m} & r_{m+1,m+1} \end{pmatrix}$$

2. *逐步回归分析的变换公式*

逐步回归分析求解回归方程就是对矩阵 $\boldsymbol{R}$ 实施一系列的矩阵变换。设逐步回归已进行了 N 步，共引入了 l 个变量 $x_{k_1},x_{k_2},\cdots,x_{k_l}$。它的第 $N+1$ 步不论是引入还是剔除变量 x_{k_α}，都是按式(3-16)对矩阵中的元素进行一次变换，得第 $N+1$ 步的矩阵。

$$r_{ij}^{(N+1)}=\begin{cases} r_{ij}^{(N)}/r_{k_\alpha k_\alpha}^{(N)} & i=k_\alpha,j\neq k_\alpha \\ r_{ij}^{(N)}-r_{ik_\alpha}^{(N)}r_{k_\alpha j}^{(N)}/r_{k_\alpha k_\alpha}^{(N)} & i\neq k_\alpha,j\neq k_\alpha \\ -r_{ij}^{(N)}/r_{k_\alpha k_\alpha}^{(N)} & i\neq k_\alpha,j=k_\alpha \\ 1/r_{k_\alpha k_\alpha}^{(N)} & i=k_\alpha,j=k_\alpha \end{cases} \tag{3-16}$$

式中 $r_{ij}^{(N)}$——$\boldsymbol{R}$ 经过第 N 步变换后的矩阵 $\boldsymbol{R}^{(N)}$ 中的元素；

$r_{ij}^{(N+1)}$——$\boldsymbol{R}$ 经过第 $N+1$ 步变换后矩阵 $\boldsymbol{R}^{(N+1)}$ 中的元素。

四、方差贡献、偏差平方和及回归系数

假设逐步回归进行了 N 步，引入了变量 $x_{k_1},x_{k_2},\cdots,x_{k_l}$，对应的回归方程为：

$$\hat{y}=b'^{(N)}_{k_1}x_{k_1}+b'^{(N)}_{k_2}x_{k_2}+\cdots+b'^{(N)}_{k_l}x_{k_l} \tag{3-17}$$

1. *方差贡献*

在第 N 步的基础上，逐步回归的第 $N+1$ 步不论是引入还是剔除变量 x_{k_α} 的方差贡献均按下式计算：

$$V_{k_\alpha}^{(N)}=r_{m+1,k_\alpha}^{(N)}\cdot r_{k_\alpha,m+1}^{(N)}/r_{k_\alpha k_\alpha}^{(N)}$$

当 $V_{k_\alpha}^{(N)}>0$ 时，第 $N+1$ 步是引入变量 x_{k_α}；当 $V_{k_\alpha}^{(N)}<0$ 时，第 $N+1$ 步是剔除变量 x_{k_α}。

2. 偏差平方和与回归平方和

第 N 步回归方程式(3-17)的偏差平方和与回归平方和分别为：

$$Q_1^{(N)} = r_{m+1,m+1}^{(N)}, \qquad Q_2^{(N)} = 1 - r_{m+1,m+1}^{(N)}$$

3. 回归系数和复相关系数

回归方程式(3-17)的系数和复相关系数为：

$$b'^{(N)}_{k_i} = r^{(N)}_{k_i,m+1} \quad (i = 1,2,\cdots,l)$$

$$R = (1 - r^{(N)}_{m+1,m+1})^{1/2}$$

最后指出，逐步回归分析不仅能够从拟定的变量中“筛选”作用大的变量建立回归方程，而且能够辅助确定变量间的相关形式。

五、逐步回归分析流程

总结逐步回归分析的计算过程，给出计算流程图(图 3-3)。

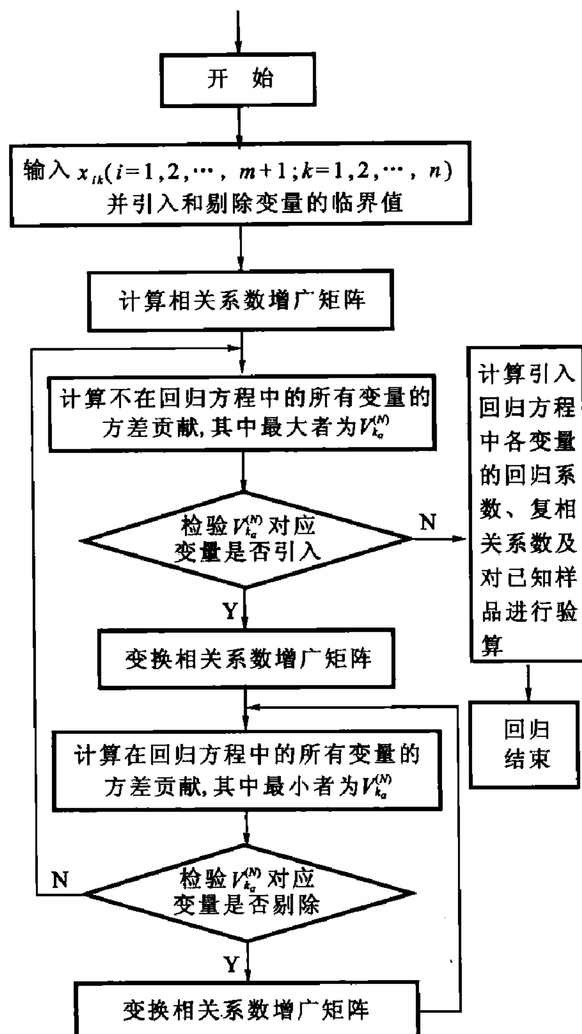


图 3-3 逐步回归分析流程图

§4 应用实例

【例 1】有利含油气面积预测。

(1)生油门限时间法。

据表 3-1 中盆地生油层数据，以埋藏深度 H、生油层温度 T 为自变量，以生油门限时间的自然对数 $\ln t$ 为因变量，求得如下回归方程：

$$\ln t = 7\,585/(T+273) - 2\,370/H - 15.0$$

该回归方程的复相关系数等于 0.94，即回归方程显著。在已知生油层温度和埋藏深度的条件下，利用上述回归方程可以预测生油层温度为 T_i、埋藏深度为 H_i 的生油门限时间 t_i。因此，生油门限时间回归方程给我们提供了一种追踪生油面积的找油途径，也就是说，用生油层埋藏时间与生油门限时间回归方程预测的生油时间之差绘制偏差等值线图，图上的正偏差区意味着有机成熟区，它是有利的勘探地区。

(2) 含油面积系数法。

陈立平、陈子恩等利用构造因素、沉积因素、生油因素对含油面积系数（单元中累计产油超过 1 t的油井数与总井数之比）进行回归分析，并利用得到的回归方程预测了四川盆地侏罗系自流井群大安寨组的含油面积系数，绘制了含油面积系数预测图（图 3-4，图中数字为百分数）。在此基础上确定了有利含油面积，为油气资源评价提供了重要依据。

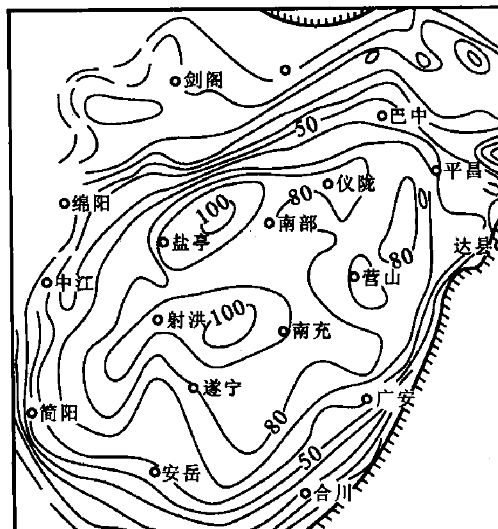


图 3-4 四川盆地侏罗系自流井群大安寨组含油面积系数预测图

基本思路：把评价区划分为 10 km×10 km 的 675 个单元，统计每个单元上反映沉积、构造、生油三个方面的 15 项地质参数及含油面积系数。675 个单元中的 139 个单元有钻探资料，另外还有 11 个边界控制单元，共计有 150 个单元。根据 150 个单元的数据，取引入与剔除变量的临界值均为 2.5，建立地质参数对含油面积系数的回归方程。

① 地质参数。

构造条件：

- x_1——大安寨组底面构造七次趋势剩余值，m；
- x_2——早第三纪前大安寨组底面古构造六次趋势剩余值，m；
- x_3——大安寨组底面现今构造海拔高度，m；
- x_4——早第三纪前大安寨组底面古构造六次趋势值，m。

沉积条件：

- x_5——介屑灰岩、页岩沉积韵律数；
- x_6——页岩厚度/介屑灰岩厚度，%；
- x_7——（页岩厚度＋介屑灰岩厚度）/组厚度，%；
- x_8——（页岩厚度/组厚度）×（介屑灰岩厚度/组厚度）（小数）；
- x_{12}——页岩厚度，m；
- x_{13}——介屑灰岩厚度，m；
- x_{14}——介屑灰岩单层平均厚度，m。

生油条件 $\begin{cases} x_9\text{——有机质成熟度时间温度指数}(TTI); \\ x_{10}\text{——单位面积生油量,t/km}^2; \\ x_{11}\text{——单位体积生油量,t/km}^3; \\ y\text{——含油面积系数,\%。} \end{cases}$

② 建立回归方程。

进行逐步回归分析，共引入七个变量，得到回归方程：

$$\hat{y} = 10^{-4}(2.64x_1 + 6.82x_2 - 2.08x_3 - 3.40x_4) + 10^{-3}(7.14x_5 + 8.442x_7) + 10^{-2} \times 2.861x_6 + 0.781$$

引入参数的组合反映大安寨组具有受构造、岩性、岩相控制的裂缝性油藏的特点。

【例 2】油气资源量预测。

(1) 油气地质条件法。

勘探实践表明，含油气凹陷中单位面积的油气储量与油气地质条件(生、储、盖、圈、保)有着密切的关系。1985 年 2 月，朱子仁等利用我国东部地区一些勘探程度较高的含油气凹陷的实际资料，建立的油气地质条件对探明储量的回归方程为：

$$y = 0.136x_1 + 0.729x_2 + 0.356x_3 + 0.152x_4 - 0.12N - 5.37$$

式中 y—— 单位面积的油气储量，10^4 t/km^2；

x_1——生油岩体积与沉积岩体积之比(生油条件)，%；

x_2——有机质转化率(生油条件)，%；

x_3—— 储集岩体积与沉积岩体积之比(储、盖条件)，%；

x_4——近油源圈闭面积与沉积岩面积之比(圈闭条件)，%；

N——含油气凹陷所经历的剥蚀次数(保存条件)。

油气储量为中值，其置信区间为 $y \pm 2S$。

采用概算储量得到回归方程：

$$y = 0.835x_1 + 0.597x_2 + 0.269x_3 + 0.142x_4 - 0.05N - 6.654$$

利用概算储量回归方程，预测前梨园洼陷沙二下段 460 km^2 内的石油资源总量为：

$$\sum Q = 28.834 \times 460 = 13\,263.64(10^4\ \text{t})$$

(2) 石油资源丰度法。

回归方程的内涵取决于拟定的地质变量。若把影响石油资源丰度的主要地质因素拟定为烃源岩的生烃强度、储层的发育程度、烃源岩上覆地层区域不整合面个数和评价单元的圈闭面积系数等，根据实际资料建立的地质因素对石油资源丰度的回归方程(据赵文智主编《石油地质理论与方法进展》,2006)为：

$$y = 0.042x_1 - 9.369x_2 + 0.297x_3 + 2.910e^{-0.434\,9x_4} - 5.688$$

式中 y——石油资源丰度，10^4 t/km^2；

x_1——烃源岩生烃强度，10^4 t/km^2；

x_2——储层厚度与沉积岩厚度之比，%；

x_3——圈闭面积系数，%；

x_4——不整合面个数。

(3) 体积速度法。

1975 年 И. И. Несмеров 根据世界 22 个勘探程度较高的含油气盆地的资料，统计得出油气总资源量(换算成油的地质储量)与盆地沉积速度的关系为：

$$\lg Q = 2.813 + 1.6131 \lg V$$

式中 Q——油气地质储量，$\times 10^6$ t；

V——沉积物充填的平均体积速度，10^3 km^3/Ma。

22 个盆地分为四类(图 3-5)：

Ⅰ类：波斯湾、墨西哥湾、西西伯利亚等，其平均体积速度大于 14×10^3 km^3/Ma；

Ⅱ类：伏尔加-乌拉尔、马拉开波、南里海盆地等，其平均体积速度为$(4\sim14)\times10^3$ km^3/Ma；

Ⅲ类：二叠盆地、圣华金盆地、切尔斯-克里海等，其平均体积速度为$(1.5\sim4)\times10^3$ km^3/Ma；

Ⅳ类：多是些小盆地，如维也纳、伊里诺斯、密执安盆地、亚速夫-库班、丹佛、泡德河、维灵斯顿等盆地，其平均体积速度小于 1.5×10^3 km^3/Ma。

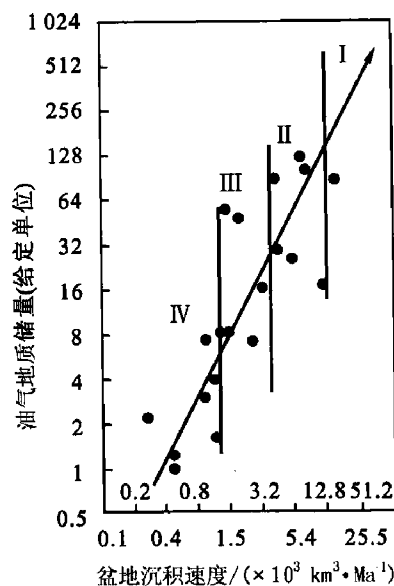


图 3-5 油气原始地质储量与沉积充填平均体积速度

【例 3】参数预测。

(1)油气运移聚集系数预测。

油气运移聚集系数是成因法估算油气资源量的关键参数。在中国石油第三次资源评价中，对 38 个资源探明程度相对较高的油气聚集单元进行剖析，获得了油气成藏条件定量描述参数和油气运移聚集系数等重要参数。在此基础上分析、研究了油气成藏地质因素与油气运移聚集系数的关系，采用逐步回归分析方法，建立的油气运移聚集系数预测模型(据赵文智主编《石油地质理论与方法进展》，2006)为：

$$\ln y = 1.478 - 0.00318x_1 + 0.186x_2 - 0.112x_3 + 0.02118x_4$$

式中 y——石油运移聚集系数，%；

x_1——烃源岩年龄，Ma；

x_2——烃源岩成熟度，%；

x_3——不整合面个数；

x_4——圈闭面积系数，%。

(2) 测井盖层突破压力预测。

泥质盖层封闭能力主要反映在两个方面：一是微观封闭能力的强弱；二是宏观展布范围。根据盖层阻止油气运移的方式，可把盖层的微观封闭机理分为毛细管力封闭、异常压力封闭和浓度封闭。对毛细管力封闭机理而言，泥质盖层的突破压力是评价的关键参数。测井资料计算突破压力是通过测井计算的总孔隙度和有效孔隙度进行的。

由苏丹 Muglad 盆地岩心与测井资料回归的突破压力方程(据方朝亮主编《勘探开发集成配套技术及应用实践》，2006)为：

$$p_{a_1} = e^{-k_1}\varphi_t$$

$$p_{a_2} = -200\varphi_e + 224.36$$

式中 p_{a_1}——总孔隙度计算的泥质盖层突破压力，MPa；

φ_t——泥质盖层的总孔隙度，%；

k_1——经验系数，当 $\varphi_t>20\%$时 $k_1=-0.2$，当 $\varphi_t<20\%$时，$k_1=-0.22$；

p_{a_2}——有效孔隙度计算的泥质盖层突破压力，MPa；

φ_e——泥质盖层的有效孔隙度，%。

思考与练习

1. 简述回归分析的基本概念和研究对象。
2. 什么是多元线性回归模型、回归方程？如何拟定回归模型？
3. 简述最小二乘法确定回归方程系数的基本原理。
4. 简要归纳多元线性回归分析的计算步骤。
5. 简述回归方程显著性的检验方法及其含义。
6. 如何求非线性回归方程？
7. 简述逐步回归分析的基本思想。
8. 结合专业知识总结回归分析在地质研究和油气勘探中的应用。

9. 镜质体反射率 R_o 是衡量生油岩成熟度的一个常用指标，其观测值的平均值 $\bar{R}_o$ 与生油岩的埋藏深度 H 之间有

$$\bar{R}_o = a_0 + a_1 H + \varepsilon$$

的关系。现有沾化凹陷的一批 H 以及与之对应的 $\bar{R}_o$ 的观测值数据（表 3-2），试求 H 对 $\bar{R}_o$ 的回归方程，并对回归方程进行显著性检验。

表 3-2 沾化凹陷镜质体反射率平均值与生油岩埋藏深度数据

序 号	深度/m	反射率平均值	序 号	深度/m	反射率平均值	序 号	深度/m	反射率平均值
1	1 200	0.30	11	2 200	0.46	21	3 200	0.52
2	1 300	0.32	12	2 300	0.47	22	3 300	0.55
3	1 400	0.33	13	2 400	0.48	23	3 400	0.58
4	1 500	0.33	14	2 500	0.49	24	3 500	0.62
5	1 600	0.34	15	2 600	0.49	25	3 600	0.67
6	1 700	0.35	16	2 700	0.49	26	3 700	0.70
7	1 800	0.37	17	2 800	0.50	26	3 800	0.75
8	1 900	0.38	18	2 900	0.50	28	3 900	0.79
9	2 000	0.41	19	3 000	0.50	29	4 000	0.83
10	2 100	0.44	20	3 100	0.51	30	4 100	0.87

第四章　聚类分析

§1　聚类分析与聚类统计量

一、聚类分析

地质学中有很多分类研究的问题，如沉积岩、古生物、矿物、油气藏、油气地球化学勘探指标的分类是一些直接分类的例子；油气资源评价、油源对比等是间接分类研究的问题；地层划分属于另一种分类。为了叙述方便，在此将分类的具体目标统称为样品（或变量）。聚类分析是根据样品（或变量）之间的亲疏程度，将样品（或变量）进行逐级定量分类的一种多元统计分析方法。根据分类的基本原理，聚类分析可分为聚合法和分解法。

1. 聚合法聚类分析

某地区油气地表化探样品具有 14 项指标，按指标间的相关程度分类，若取相关系数为 0.75，可把指标分为三类，最终以 0.07 的相关程度合并为一类（图 4-1）。在此基础上给出聚合法聚类分析的一般概念。

聚合法聚类分析又称为点群分析，它是按样品（或变量）在性质（成因）上的亲疏关系，把样品（或变量）进行逐级定量分类的一种多元统计分析方法。这种聚类分析，开始时每个样品（或变量）自成一类，然后以某种表示样品（或变量）间亲疏关系的统计量（分类指标）为分类依据，先把彼此关系最密切的样品（或变量）合并为一类，再把关系相对亲近的小类（一个或多个样品（或变量））合并成一类……，直到所有的样品（或变量）合并成一个大类为止。最终的分类结果是一幅反映样品（或变量）间的差异和亲疏关系（由细分类到粗分类）的定量分类系统图——聚类分析谱系图（图 4-1）。

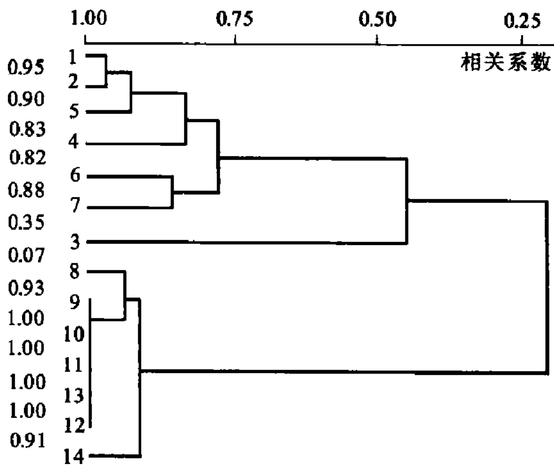


图 4-1　油气地球化学勘探指标聚类谱系图
1,2,…,14 为修整指标编号

研究样品的相似性，对样品的分类又称为 Q 型聚类分析；而研究变量的相关性，对变量的分类则叫做 R 型聚类分析。

聚合法聚类分析中的样品（或变量），如油气藏、沉积岩、地球化学勘探指标之间没有次序约束关系。

2. 分解法聚类分析

从分类原理上讲，这种聚类分析与聚合法聚类分析恰好相反，开始时全部样品为一个大类，依据某种分类指标，把全部样品分为两类、三类……，直到满足分类的要求为止。例如，假设数列{1,1,2,2,3,3}是六个样品某个变量的观测值，根据观测值的相似性，把六个样品分为三类，结果是{1,2}，{3,4}，{5,6}。因此，分解法聚类分析是把大类分解成小类的统计

分析方法。

分解法聚类分析中的样品之间有次序约束关系，如沿着地层剖面依次取若干块岩石样品，若将样品的分类结果用于地层划分或者岩性识别，那么分类时样品的排列顺序就不能打乱。

二、聚类统计量

聚类统计量是衡量分类对象相似(相关)程度的统计指标。在此介绍几个最常用的聚类统计量。

1. 聚合法聚类统计量

假设有 n 个样品，每个样品有 m 个变量，它们的观测值 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$ 构成一个数据矩阵，记为：

$$\boldsymbol{X}_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{4-1}$$

由式(4-1)可以看出：

① 矩阵的行元素是样品的观测值，可将其视为 m 维空间的一个点或一个矢量。

② 矩阵的列元素是变量的 n 次观测值，可将其视为 n 维空间的一个点或一个矢量。

③ 由上可知，样品间的相似性就是数据矩阵中行之间的相似性，Q 型聚类就是把数据矩阵中相似程度高的行合并为同类；变量间的相关性就是数据矩阵中列之间的相关性，R 型聚类就是把数据矩阵中相关程度高的列合并为同类。在此基础上，给出聚合法聚类统计量。

(1) Q 型聚类分析统计量。

① 相似系数。

设矢量 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{im})$，$\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jm})$，定义 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 夹角 θ_{ij} 的余弦为 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 的相似系数，即：

$$r_{ij}=\cos\theta_{ij}=\frac{\boldsymbol{X}_i\cdot\boldsymbol{X}_j}{|\boldsymbol{X}_i||\boldsymbol{X}_j|}=\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\Big/\left(\sqrt{\sum_{k=1}^{m}x_{ik}^2}\sqrt{\sum_{k=1}^{m}x_{jk}^2}\right)\quad(i,j=1,2,\cdots,n) \tag{4-2}$$

在矩阵 $(r_{ij})_{n\times n}$ 中，$r_{ij}=r_{ji}$，$r_{ii}=1$。r_{ij} 越接近于 1，$\boldsymbol{X}_i$，$\boldsymbol{X}_j$ 的性质越相近。

② 相关系数。

定义矢量 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 的相关系数为：

$$r_{ij}=\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)\Big/\left(\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2}\right)\quad(i,j=1,2,\cdots,n) \tag{4-3}$$

在矩阵 $(r_{ij})_{n\times n}$ 中，$r_{ij}=r_{ji}$，$r_{ii}=1$ 。r_{ij} 越接近于 1，$\boldsymbol{X}_i$，$\boldsymbol{X}_j$ 的性质越相近。

③ 距离系数。

在 m 维空间中，两个点之间的距离为：

$$d_{ij}=\left[\sum_{k=1}^{m}(x_{ik}-x_{jk})^2\right]^{1/2}\quad(i,j=1,2,\cdots,n)$$

为了防止 d_{ij} 过大而造成计算溢出，把上式改写为：

$$d_{ij} = \left[\frac{1}{m}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2\right]^{1/2} \quad (i,j = 1,2,\cdots,n) \tag{4-4}$$

在矩阵 $(d_{ij})_{n\times n}$ 中，$d_{ij}=d_{ji}$，$d_{ii}=0$ 。d_{ij} 越接近于 0，$\boldsymbol{X}_i$，$\boldsymbol{X}_j$ 的性质越相近。

(2) R 型聚类分析统计量。

如前所述，变量间的相关性是数据矩阵中列之间的相关性。因此，仿照 Q 型聚类统计量，并设矢量 $\boldsymbol{X}_i=(x_{1i},x_{2i},\cdots,x_{ni})$，$\boldsymbol{X}_j=(x_{1j},x_{2j},\cdots,x_{nj})$，易写出 R 型聚类分析统计量：

① 相似系数。

$$r_{ij} = \cos\ \theta_{ij} = \frac{\boldsymbol{X}_i \cdot \boldsymbol{X}_j}{|\boldsymbol{X}_i||\boldsymbol{X}_j|} = \sum_{k=1}^{n}x_{ki}\cdot x_{kj}\Big/\left(\sqrt{\sum_{k=1}^{n}x_{ki}^2}\sqrt{\sum_{k=1}^{n}x_{kj}^2}\right) \quad (i,j = 1,2,\cdots,m) \tag{4-5}$$

在矩阵 $(r_{ij})_{m\times m}$ 中，$r_{ij}=r_{ji}$，$r_{ii}=1$ 。r_{ij} 越接近于 1，$\boldsymbol{X}_i$，$\boldsymbol{X}_j$ 的相关越密切。

② 相关系数。

$$r_{ij} = \sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)\Big/\left(\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}\right) \quad (i,j = 1,2,\cdots,m) \tag{4-6}$$

在矩阵 $(r_{ij})_{m\times m}$ 中，$r_{ij}=r_{ji}$，$r_{ii}=1$ 。r_{ij} 越接近于 1，X_i 与 X_j 相关越密切。

③ 距离系数。

$$d_{ij} = \left[\frac{1}{n}\sum_{k=1}^{n}(x_{ki} - x_{kj})^2\right]^{1/2} \quad (i,j = 1,2,\cdots,m) \tag{4-7}$$

在矩阵 $(d_{ij})_{m\times m}$ 中，$d_{ij}=d_{ji}$，$d_{ii}=0$。d_{ij} 越接近于 0，$\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 相关越密切。

2. 分解法聚类统计量

设有 n 个样品，每个样品有 m 个变量，它们的观测值 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$ 构成数据矩阵式(4-1)。

分解法聚类分析的实质是找出数据矩阵中数据的变化界线，把数据矩阵分段。若把它分成 k 段，每段内有 $n_l(l=1,2,\cdots,k)$ 个样品，分段结果记为：

$$\underset{\text{第 1 段}}{\begin{pmatrix} x_{11}^{(1)} & x_{11}^{(2)} & \cdots & x_{11}^{(m)} \\ x_{12}^{(1)} & x_{12}^{(2)} & \cdots & x_{12}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{1n_1}^{(1)} & x_{1n_1}^{(2)} & \cdots & x_{1n_1}^{(m)} \end{pmatrix}} \quad \underset{\text{第 2 段}}{\begin{pmatrix} x_{21}^{(1)} & x_{21}^{(2)} & \cdots & x_{21}^{(m)} \\ x_{22}^{(1)} & x_{22}^{(2)} & \cdots & x_{22}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{2n_2}^{(1)} & x_{2n_2}^{(2)} & \cdots & x_{2n_2}^{(m)} \end{pmatrix}} \cdots \underset{\text{第 } k \text{ 段}}{\begin{pmatrix} x_{k1}^{(1)} & x_{k1}^{(2)} & \cdots & x_{k1}^{(m)} \\ x_{k2}^{(1)} & x_{k2}^{(2)} & \cdots & x_{k2}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{kn_k}^{(1)} & x_{kn_k}^{(2)} & \cdots & x_{kn_k}^{(m)} \end{pmatrix}}$$

把数据矩阵分为 k 段，有不同的分法。对于其中的任何一种分法：

总离差平方和：

$$S = \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{ij} - \bar{x}_l^{(i)})^2 \quad \text{（为一常数）} \tag{4-8}$$

段内离差平方和：

$$S_1 = \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{ij} - \bar{x}_l^{(i)})^2 \tag{4-9}$$

段间离差平方和：

$$S_2 = \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(\bar{x}_l^{(i)} - \bar{x}^{(i)})^2 = \sum_{l=1}^{k} n_l \sum_{i=1}^{m}(\bar{x}_l^{(i)} - \bar{x}^{(i)})^2 \qquad (4\text{-}10)$$

式中　n——样品的总数，$n = \sum_{l=1}^{k} n_l$；

$\bar{x}_l^{(i)}$——第 l 段内第 i 个变量 n_l 个观测值的平均值，$\bar{x}_l^{(i)} = \frac{1}{n_l}\sum_{j=1}^{n_l} x_{lj}^{(i)}$；

$\bar{x}^{(i)}$——第 i 个变量 n 个观测值的平均值，$\bar{x}^{(i)} = \frac{1}{n}\sum_{l=1}^{k}\sum_{j=1}^{n_l} x_{lj}^{(i)} = \frac{1}{n}\sum_{l=1}^{k} n_l \bar{x}_l^{(i)}$；

$x_{lj}^{(i)}$——第 l 段内第 j 个样品第 i 个变量的观测值。

可以证明：

$$S = S_1 + S_2$$

对于给定的数据矩阵来说，S 是一个常数，故 S_1 小，S_2 就大。因此，可以把 S_1 作为分解法聚类分析的聚类统计量。它的内涵是：S_1 越小，各段内数据的离差越小（各段内样品的差异越小），而各段之间数据的离差越大（各段之间样品的差异越大）。

§2　聚合法聚类分析

如前所述，聚合法聚类分析是将样品（或变量）的类由多变少，直到把全部样品合并成一类的定量分类方法，其聚类过程大致如下：

① 原始类，每个样品（或变量）自成一类，这时类的数目等于样品（或变量）的个数。

② 第一级聚类。据某一种统计量计算各个样品（或变量）之间的亲疏程度，把关系密切的样品（或变量）合并成一类，并将该类改造成一个代表性样品（或变量）参加下一级聚类。

③ 第二级聚类。再计算样品（或变量）间的亲疏程度，把关系密切的样品（或变量）合并为一类（样品（或变量）与样品（或变量）或样品（或变量）与小类合并），并将该类改造成一个代表性样品（或变量）参加下一级聚类。

④ 按上述方法进行第三级聚类、第四级聚类……，直到全部样品（或变量）合并为一个大类为止。

在上述聚类过程中，要不断地计算样品（或变量）与样品（或变量）、样品（或变量）与类、类与类之间的亲疏程度，下面介绍计算它们之间亲疏程度的递推公式和聚类过程。

一、距离类统计量聚合法

对于 m 维或 n 维空间的两个点来说，两点间的距离是两点间线段的长度。但是，两类间的距离就有不同的定义，既可把两类中相距最近的两个点，又可把两类中相距最远的两个点之间的距离作为两类间的距离，还可以取两类的重心距离作为两类间的距离等（图 4-2）。因距离的定义不同，也就产生了不同的聚合法。下面介绍四种距离度量聚合法。

1. 最短距离法

（1）最短距离。

如果类 p 与类 q 合并为类 r，那么定义类 p、类 q 间的最短距离为：

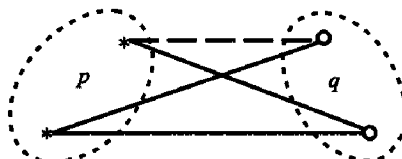


图 4-2　类间距离示意图

$$D_{pq}=\min_{\substack{x_i\in p\\x_j\in q}} d_{ij} \tag{4-11}$$

式中　d_{ij}——类 p、类 q 中的样品（或变量）点 x_i 与 x_j 间的距离。

最短距离下的合并分类叫做最短距离法聚类。计算新的类 r 与某类 f 之间最短距离的递推公式为：

$$D_{rf}=\frac{1}{2}(D_{fp}+D_{fq})-\frac{1}{2}\mid D_{fp}-D_{fq}\mid \tag{4-12}$$

(2) 最短距离法聚类的基本过程。

① 准备工作。

计算初始的距离矩阵 $D^{(0)}$。对于样品，$D^{(0)}=(d_{ij}^{(0)})_{n\times n}$；对于变量，$D^{(0)}=(d_{ij}^{(0)})_{m\times m}$。

② 第一级聚类。

把 $D^{(0)}$ 中最短距离对应的类合并为类 r。按式(4-12)计算类 r 与其他各类的最短距离矩阵，记为 $D^{(1)}$。

③ 第二级聚类。

把 $D^{(1)}$ 中最短距离对应的类合并为类 k。按式(4-12)计算类 k 与其他各类的最短距离矩阵，记为 $D^{(2)}$。

⋮

重复上述计算过程，直到全部样品（或变量）聚为一个大类为止。

2. 最长距离法

(1) 最长距离。

如果类 p 与类 q 合并为类 r，那么定义类 p、类 q 间的最长距离为：

$$D_{pq}=\max_{\substack{x_i\in p\\x_j\in q}} d_{ij} \tag{4-13}$$

式中　d_{ij}——类 p、类 q 中的样品（或变量）点 x_i 与 x_j 间的距离。

最长距离下的合并分类叫做最长距离法聚类。计算新的类 r 与某类 f 之间最长距离的递推公式为：

$$D_{rf}=\frac{1}{2}(D_{fp}+D_{fq})+\frac{1}{2}\mid D_{fp}-D_{fq}\mid \tag{4-14}$$

(2) 最长距离法聚类的基本过程。

最长距离法聚类的基本过程与最短距离法基本相同，不同之处是按式(4-14)计算最长距离矩阵，并按最长距离合并类。

3. 类平均法

(1) 类平均距离。

如果类 p 与类 q 合并为类 r，那么定义类 p、类 q 间的平均距离为：

$$D_{pq}=\frac{1}{n_p n_q}\sum_{\substack{x_i\in p\\x_j\in q}} d_{ij} \tag{4-15}$$

式中　d_{ij}——类 p、类 q 中的样品（或变量）点 x_i 与 x_j 间的距离；

n_p,n_q——类 p、类 q 中的样品（或变量）数，类 r 中的样品（或变量）数 $n_r=n_p+n_q$。

在类平均距离下进行的聚合聚类叫做类平均法。计算新类 r 与某类 f 之间平均距离的

递推公式为：

$$D_{rf}=\frac{n_p}{n_r}D_{fp}+\frac{n_q}{n_r}D_{fq} \tag{4-16}$$

(2) 类平均法聚类的基本过程。

类平均法聚类的基本过程与上述聚类过程类似，不同之处是计算类平均距离矩阵。

4. 重心法

假设类 p、类 q 的重心分别是 $\bar{X}_p,\bar{X}_q$，那么类 p 与类 q 间的重心距离为：

$$D_{pq}=d_{\bar{X}_p\bar{X}_q} \tag{4-17}$$

类 p 与类 q 合成新类 r 后，新类 r 的重心为：

$$\bar{X}_r=\frac{1}{n_r}(n_p\bar{X}_p+n_q\bar{X}_q)$$

计算新类 r 与某类 f 之间重心距离的递推公式为：

$$D_{rf}=\frac{n_p}{n_r}D_{fp}^2+\frac{n_q}{n_r}D_{fq}^2-\frac{n_pn_q}{n_r^2}D_{pq}^2 \tag{4-18}$$

用重心距离进行的聚合聚类叫做重心法聚类。

二、相关类统计量聚合法

相关类统计量包括相关系数、相似系数。相应的聚合法有近邻连接法、远邻连接法和类平均法。计算新类 r 与某类 f 之间相关类统计量的相应递推公式为：

1. 近邻连接法

$$R_{rf}=\frac{1}{2}(R_{fp}+R_{fq})+\frac{1}{2}|R_{fp}-R_{fq}| \tag{4-19}$$

2. 远邻连接法

$$R_{rf}=\frac{1}{2}(R_{fp}+R_{fq})-\frac{1}{2}|R_{fp}-R_{fq}| \tag{4-20}$$

3. 类平均法

$$R_{rf}=\frac{n_p}{n_r}R_{fp}+\frac{n_q}{n_r}R_{fq} \tag{4-21}$$

三、聚类结果的选择

不同的系统聚类方法得到的结果有差异，如云南省某地区超基性岩体岩样的聚类分析结果(图 4-3)就不一样。究竟哪种方法的分类结果好，目前尚无合适的衡量标准。在实际应用中，要结合其他地质理论及资料，分析不同方法给出的分类结果，从中确定一种合理的分类方案。

四、数据预处理

由于变量的单位和观测值的数量级可能不同，若直接用原始观测值计算聚类统计量，就会突出观测值绝对值大的变量的作用，压低观测值绝对值小的变量的影响。因此，在进行聚类分析时应采用合适的数据处理方法(参见第二章)对原始数据进行标准化处理。

五、聚类分析流程

总结上述聚类分析过程，给出如下聚类分析流程图(图 4-4)。

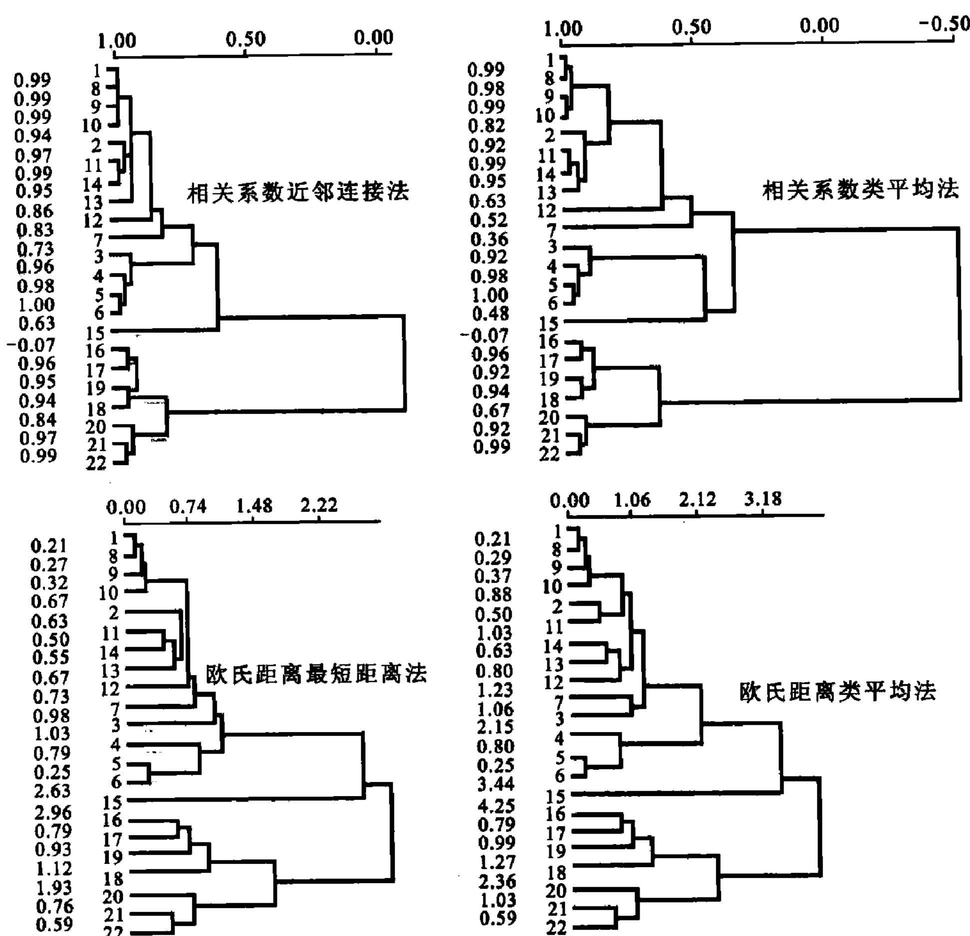


图 4-3　超基性岩体岩样聚类谱系图

§3　分解法聚类分析

一、分解法聚类的基本思想

如果把数据矩阵式(4-1)视为一个数据序列，那么分解法聚类分析就是以段内离差平方和为分类统计量，找出数据序列的变化界线，把数据序列分段的统计分析方法。把数据序列分段后，若满足段内离差平方和最小，则是最优的分类方案。这种分类方案又叫做最优分割。

最优 k 分割的基本思想是按照段内离差平方和最小的原则，依次找出数据序列的 $k-1$ 个分点，把数据序列分为 k 段。

二、分解法聚类的基本过程

为了书写方便，在此把数据序列简记为：

$$X_{n\times m} = (X_1 X_2 \cdots X_n)'$$

式中　$X_i=(x_{i1},x_{i2},\cdots,x_{im})(i=1,2,\cdots,n)$。

1. 符号说明

符号 $S_n(k;j)$是把 $X_{n\times m}$分为 k 段的段内离差平方和，n 是 $X_{n\times m}$中的样品数，k 是分段数，$j(1\leqslant j\leqslant n-1)$表示以第 j 个样品后为分界点。

2. 分段过程

(1) 分 2 段。

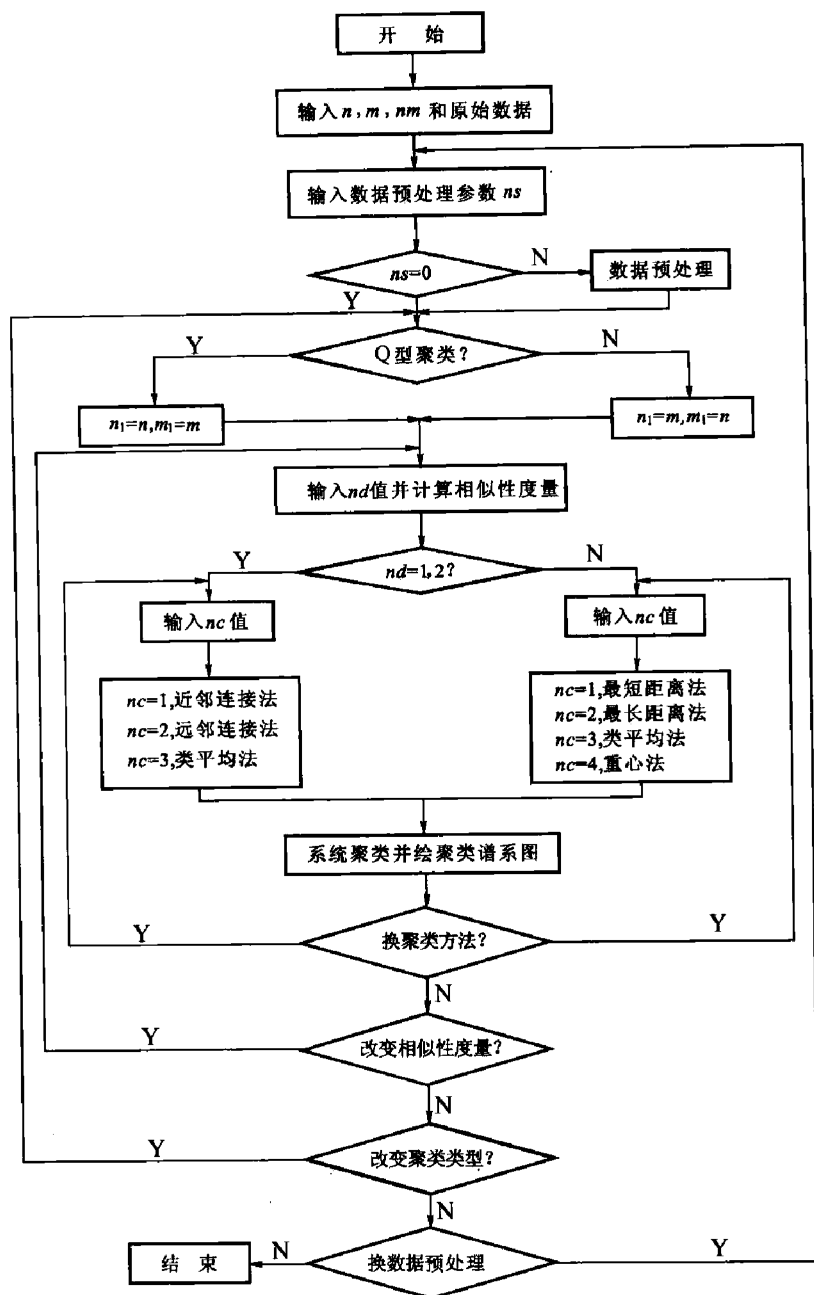


图 4-4 聚合法聚类分析流程图

当 $k=2$ 时,令 $j=1,2,\cdots,n-1$,按式(4-9)计算 $S_n(2;j)$。若 $S_n(2;\alpha_1)=\min\limits_{1\leqslant j\leqslant n-1} S_n(2;j)$,那么,$\alpha_1$ 是分 2 段的最优分界点,分段结果如下:

$$[(\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_{\alpha_1})(\boldsymbol{X}_{\alpha_1+1}\cdots\boldsymbol{X}_n)]'$$

(2) 分 3 段。

当 $k=3$ 时,令 $j=1,2,\cdots,n-1,j\neq\alpha_1$,按式(4-9)计算 $S_n(3;j)$,若 $S_n(3;\alpha_2)=\min\limits_{1\leqslant j\leqslant n-2} S_n(3;j)$,那么 α_2 是分 3 段的最优的第 2 个分界点,分段结果如下:

$$[(\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_{\alpha_1})(\boldsymbol{X}_{\alpha_1+1}\cdots\boldsymbol{X}_{\alpha_2})(\boldsymbol{X}_{\alpha_2+1}\cdots\boldsymbol{X}_n)]'$$

⋮

(3) 分 k 段。

当 $k=h$ 时，令 $j=1,2,\cdots,n-1,j\neq\alpha_1,\alpha_2,\cdots,\alpha_{k-2}$，按式(4-9)计算 $S_n(h;j)$，若 $S_n(h;\alpha_{k-1})=\min\limits_{1\leqslant j\leqslant n-k+1}S_n(h;j)$，那么 α_{k-1} 是分 k 段的最优的第 $k-1$ 个分界点，分段结果如下：

$$[(\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_{\alpha_1})(\boldsymbol{X}_{\alpha_1+1}\cdots\boldsymbol{X}_{\alpha_2})\cdots(\boldsymbol{X}_{\alpha_{k-1}}\cdots\boldsymbol{X}_n)]'$$

三、分段数的确定

对于分段数 k，根据数据序列的曲线形态，可以给出 k 的估计值 k_0。另外，也可以预先给定一个小正数 δ，当段内离差平方和 $S_n(k;j)<\delta$ 时结束分段，与此时的段内离差平方和对应的 k 就是最后的分段数。还可以绘制段内离差平方和随分段数的变化曲线(图 4-5)，该曲线随着分段数的增加而单调递减，并趋于平缓，可以选择曲线开始平缓时的点对应的 k 为最优的分段数。

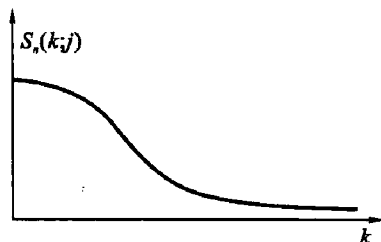


图 4-5 段内离差平方和与分段数的关系

四、最优分割计算过程

总结计算机实现最优分割的过程，给出最优分割流程图(图 4-6)。

§4 应用实例

【例 1】油气地表化探指标的分类。

在油气地表化探工作中，一般要分析样品的甲烷、乙烷、丙烷、紫外、荧光、总烃和重烃等20余项指示地下油气的指标。这些指标既有相对的独立性，又存在着一定的成因联系。因此，可利用聚类分析对它们分类，从而研究它们的成因联系，选择有代表性的指标，化简油气地表化探工作。

在中国东部某地区油气地表化探样品 29 项指标的聚类谱系图(图 4-7)上，取相关系数 $r=0.75$，把指标分成 13 类。其中成因联系密切的有酸解烃、紫外和荧光类。

酸解烃类包含了全部的酸解烃指标。其中甲烷与总烃以 $r=0.98$ 聚为一类，其原因在于甲烷占了总烃的绝大部分，并有密切的成因联系。甲烷在油气藏中的浓度最高，运移能力强，到达地表的数量最多。重烃的综合性强，易被土壤吸附，直接反映烃场浓度的大小和地下油气的性质。因此，可将它们作为油气地表化探的重要指标。

紫外和荧光类包括了紫外(紫外 216 除外)和荧光指标。它们都是检测油气藏中芳烃混合物的直接指标。因此，可从这两类中各选一个指标作为它们的代表性指标。

其他分散指标可供参考。

根据上述分析，舍弃部分指标后，既可保证油气地表化探工作成果的质量，又可减少油气地表化探的工作量。

【例 2】盆地含油气远景评价。

应用聚类分析可对不同勘探程度的含油气盆地进行远景评价，其实质是地质参数的类比。因此，评价结果的可靠性依赖于参与评价的地质参数。选取一定数量的含油气盆地(勘探程度较高的和待评价的盆地)以及盆地之间可类比的定量和定性地质参数作聚类分析，可由已知含油气远景的盆地推测待评价盆地的含油气远景。

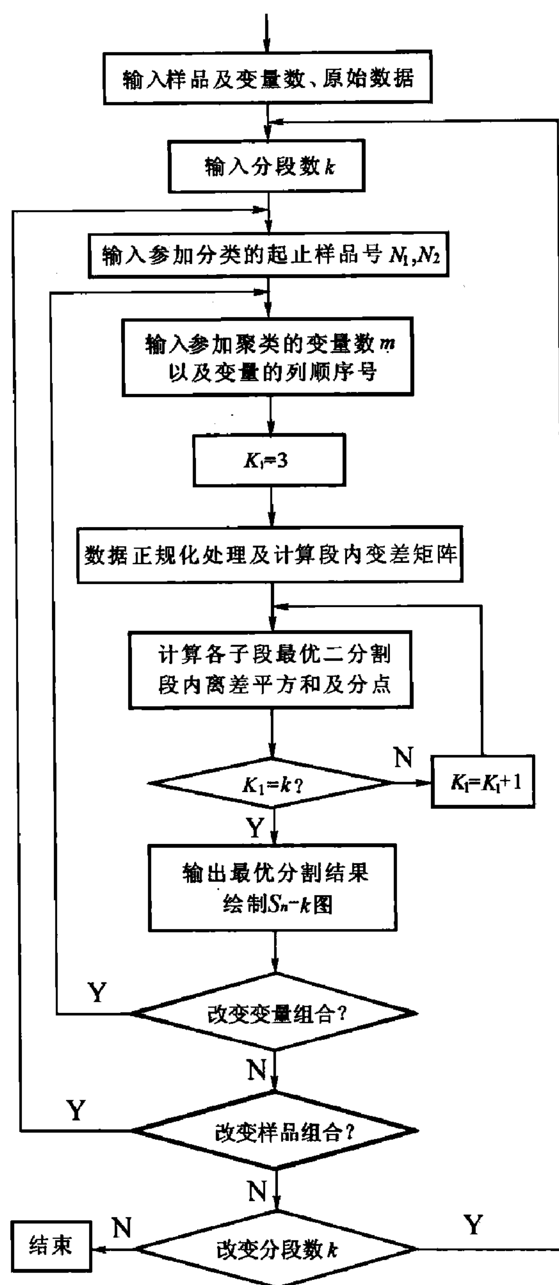


图 4-6　最优分割流程图

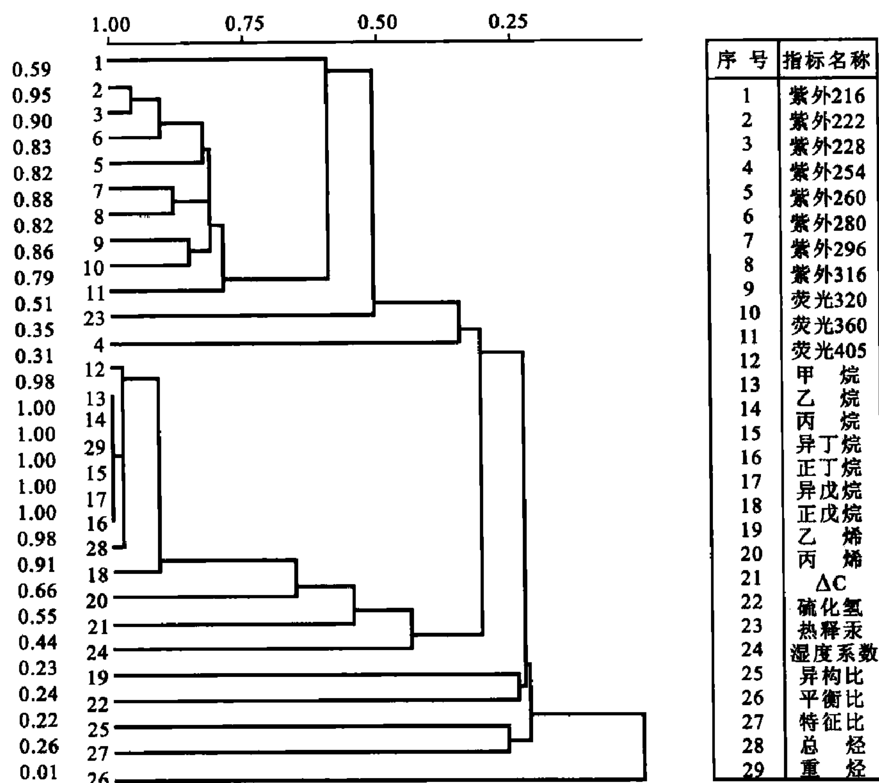
本例选取了 48 个含油气盆地，类比地质参数分 5 类 30 项，原始数据见表 4-1。对于其中的定性地质参数，如盆地类型、地质时代等，要进行定量化处理，处理方法是：

① 盆地类型：分内陆、沿海、海湾和海洋 4 项，按是为 1，不是为 0 编码。

② 地质时代：分第三纪、白垩纪、侏罗纪、三叠纪、二叠纪、石炭纪、泥盆纪、志留纪、奥陶纪和寒武纪共 10 项，按有为 1，无为 0 编码。

③ 岩性：分砂岩、碳酸盐岩、火成岩和基岩 4 项，按有为 1，无为 0 编码。

在表 4-1 中，盆地类型、沉积时代和储层时代、储层岩性分别按①，②，③中的顺序排列。盆地面积的单位为 10^5 km²，厚度的单位是 10^4 m。



序 号	指标名称
1	紫外216
2	紫外222
3	紫外228
4	紫外254
5	紫外260
6	紫外280
7	紫外296
8	紫外316
9	荧光320
10	荧光360
11	荧光405
12	甲 烷
13	乙 烷
14	丙 烷
15	异丁烷
16	正丁烷
17	异戊烷
18	正戊烷
19	乙 烯
20	丙 烯
21	ΔC
22	硫化氢
23	热释汞
24	湿度系数
25	异构比
26	平衡比
27	特征比
28	总 烃
29	重 烃

图 4-7 油气地表化探指标聚类谱系图

对原始数据进行标准差标准化后，分别采用相关系数近邻连接法和欧氏距离最短距离法进行系统聚类，得聚类谱系图(图 4-8)。因地质类比参数的局限性，故本例仅是方法上的演习。尽管如此，所得结论也能说明一些问题。

【例 3】大王北洼陷甾萜烷油源对比。

油源对比是根据地质和地球化学特征，确定石油和源岩之间成因联系的工作。它包括石油与烃源岩之间以及不同储层中石油之间的对比两个方面。众所周知，烃源岩中的干酪根在一定条件下裂解形成石油和天然气，其中一部分运移到储集层中，一部分残留在烃源岩内。因此，烃源岩中的干酪根、沥青与来自该层系的油气就有着亲缘关系，在化学组成上必然有着某种程度的相似性。同源油气化学组成的相似程度高，异源油气化学组成的相似程度低。这种相似性就是我们进行油源对比的基本依据。

油源对比的基础是受地质资料支持的化学相关性，因此对比参数的选择非常重要。目前油源对比工作中，对研究对象的化学分析可以获得大量的化学组成数据，如何合理地解释这些分析资料，得出正确的结论，仍然是油源对比工作中的难点。聚类分析能够根据样品多个参数的综合特征，确定样品之间的相似程度，是油源对比的一种可行性方法。

大王北洼陷有古近系沙四上亚段、沙三段和沙一段三套有机质丰度高、类型好的泥质烃源岩。目前已在该洼陷的沙河街组发现了三个油田。通过色-质谱分析技术对烃源岩和石油的生物标志化合物组成进行了系统分析。

表 4-1　含油气盆地类比地质参数

序　号	盆　地	面　积	沉积时代	厚　度	储层时代	储层岩性	盆地类型
1	尼日利亚沿海	1.00	1100000000	0.80	1000000000	1000	0100
2	波利尼亚克	1.10	0000011111	0.40	0000011000	1000	1000
3	三叠盆地	4.50	0111111111	0.70	0001000011	1000	0100
4	锡尔特盆地	4.00	1100000000	0.59	1100000000	1101	0100
5	苏伊士-红海	4.00	1111111000	0.40	1111111000	1100	0010
6	北　海	5.00	1111111000	0.60	1111100000	1100	0001
7	荷兰-德国盆地	0.80	1111110000	0.80	0111100000	1100	1000
8	维也纳盆地	0.70	1000000000	0.45	1000000000	1001	1000
9	前喀尔巴阡山	1.10	1111000000	0.80	1100000000	1000	1000
10	伏尔加-乌拉尔	6.90	0001111000	0.60	0000111000	1100	1000
11	伯朝拉盆地	2.00	0011111000	0.35	0000111000	1100	1000
12	南里海	2.60	1010000000	1.20	1000000000	1000	1000
13	波斯湾	32.80	1101111111	1.10	1110000000	1100	0010
14	布哈提	5.30	1110000000	0.50	0110000000	1100	1000
15	柴达木盆地	1.20	1110000000	1.00	1000000000	1000	1000
16	塔里木盆地	6.10	1111111000	1.00	1010000000	1000	1000
17	准噶尔盆地	1.60	1111110000	1.30	1001100000	1000	1000
18	吐鲁番盆地	0.50	1110000000	0.75	0010000000	0000	1000
19	陕甘宁盆地	2.90	0101111000	0.70	0011000000	1000	1000
20	酒泉盆地	0.30	1110000000	0.80	1000000000	1000	1000
21	四川盆地	2.10	0101111111	1.00	0011100000	1100	1000
22	江汉盆地	0.80	1000000000	0.80	1000000000	1000	1000
23	渤海湾	3.80	1000000000	0.50	1000000000	1001	0010
24	松辽盆地	12.70	0110000000	1.80	0100000000	1000	1000
25	西西伯利亚	33.00	1101000000	0.80	0110000000	1000	1000
26	萨哈林	0.50	1100000000	0.80	1000000000	1000	0001
27	苏门答腊	2.30	1000000000	0.75	1000000000	1100	0001
28	北爪哇盆地	0.40	1000000000	0.60	1000000000	1110	0001
29	沙捞越	0.35	1000000000	0.60	1000000000	1000	0001
30	东加里曼丹	1.50	1010000000	0.90	1000000000	1100	0001
31	佩斯盆地	1.04	1101111110	1.50	0011100000	1000	1000
32	库珀盆地	1.27	0001100000	0.37	0000100000	1000	1000
33	吉普斯兰	0.40	1100000000	0.70	1100000000	1000	0010
34	库克湾	0.39	1111000000	0.90	1000000000	1000	0010
35	阿尔伯达盆地	6.00	0111111111	0.60	0101111000	1100	1000
36	汾河盆地	0.60	1111111111	0.50	0111100000	1000	1000
37	落基山诸盆地	6.00	1111111111	1.80	1111110000	1100	1000
38	丹佛盆地	1.30	1111111111	0.50	0111100000	1000	1000
39	加里福尼亚诸盆地	1.80	1100000000	1.60	1100000000	1001	0100
40	阿巴拉契亚	4.70	0000011111	0.60	0000011111	1100	1000
41	北美地台	11.30	0000111111	1.00	0000100010	1100	1000
42	西德克萨斯(二)	3.20	0000111111	0.60	0000111110	1100	1000
43	墨西哥湾	11.00	1110000000	1.00	1110000000	1100	0010
44	维拉克鲁斯-塔巴斯克	1.60	1110000000	0.7	1100000000	1100	0010
45	马拉开波-法尔康	0.85	1100000000	0.9	1100000000	1101	0100
46	普图马约	2.60	1111000000	0.5	0100000000	1000	1000
47	圣克鲁斯	2.80	1111111000	0.6	0100011000	1000	1000
48	瓜亚及尔	0.25	1100000000	0.75	1000000000	1000	0100

原油和烃源岩甾萜烷生物标志物的组成受母源输入、沉积环境和成熟度等多种因素影响，利用质谱图鉴定的全部化合物峰面积的相对含量包含所有这些信息，直接利用聚类分析进行油源对比，可以最大限度地利用生物标志化合物的所有成因信息，减少人为选择一个或几个参数进行对比的片面性。对大王北洼陷采集的所有原油样品和烃源岩样品，以 $m/z=217$ 和 $m/z=191$ 鉴定的包括孕甾烷、升孕甾烷、重排甾烷、规则甾烷、三环萜烷、五环三萜烷在内的 53 种化合物的相对含量为变量，进行 Q 型聚类分析，得到了原始数据相关系数远邻连

1.00 0.80 0.60 0.40

尼日利亚沿海 0.99 瓜亚及尔 0.67 锡尔特盆地 0.97 马拉开波-法尔康 0.86 加里福尼亚诸盆地 0.50 维也纳盆地 0.73 渤海湾 0.62 前喀尔巴阡山 0.80 普图马约 0.63 南里海 0.76 柴达木盆地 0.98 九泉盆地 0.75 江汉盆地 0.54 吉普斯兰 0.78 墨西哥湾 0.83 维拉克鲁斯-塔巴斯克 0.65 库克湾 0.52 布哈提 0.45 萨哈林 0.85 苏门答腊 0.89 东加里曼丹 0.88 沙捞越 0.54 北爪哇盆地 0.43 波利尼亚克 0.63 阿尔伯达盆地 0.61 伏尔加-乌拉尔 0.91 伯朝拉盆地 0.60 阿巴拉契亚 0.78 西德克萨斯(二) 0.64 北美地台 0.58 圣克鲁盆地 0.55 陕甘宁盆地 0.58 四川盆地 0.69 佩斯盆地 0.64 汾河盆地 0.99 丹佛盆地 0.55 落基山诸盆地 0.55 库珀盆地 0.48 塔里木盆地 0.47 三叠盆地 0.46 北海 0.57 荷兰-德国盆地 0.50 准噶尔盆地 0.44 苏伊士-红海 0.41 波斯湾 0.48 西西伯利亚 0.38 松辽盆地 0.23 吐鲁番盆地

(a) 相关系数近邻连接法聚类谱系图

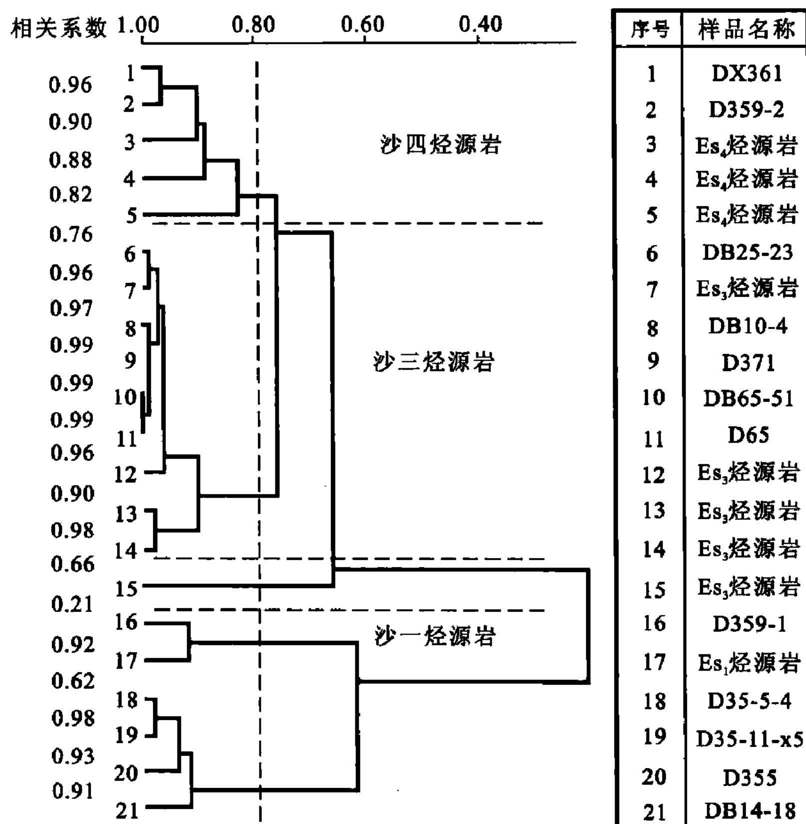
1.00 0.80 0.60 0.40

1 0.18 48 4.12 8 3.39 9 2.18 46 2.82 12 2.22 15 0.59 20 2.30 22 3.49 23 3.56 14 3.65 26 2.21 27 2.02 30 2.07 29 3.99 33 2.82 43 2.69 44 3.49 34 4.31 4 1.25 45 2.85 39 4.59 7 4.32 17 4.43 16 4.39 19 4.63 24 4.67 6 4.82 32 4.98 21 4.11 31 4.39 36 0.10 38 5.06 10 2.22 11 4.78 47 5.25 35 5.39 37 5.44 25 5.53 5 5.74 41 5.79 2 6.38 40 5.75 42 6.82 13 6.94 28 7.52 18 8.09 3

(b) 欧氏距离最短距离法聚类谱系图
(序号1~48为盆地编号,其对应的盆地名称见表4-1)

图 4-8 含油气盆地聚类谱系图

接法聚类谱系图(图 4-9)。从谱系图上可以看到,若以相关系数 $r=0.8$ 作为分类标准,研究区的原油和烃源岩可分为五类:DX361 和 D359-2 原油与沙四段烃源岩为一类,油、源间的相关系数在 0.82～0.90 之间,表明这两个原油样品的成因与沙四段烃源岩联系最密切;DB25-23,DB10-4,D371,D65-51 和 D65 原油样品与所有埋藏在 3 000～3 700 m 深度范围内的沙三段烃源岩为一类,油、源间的相关系数都在 0.90 以上,表明它们的成因与沙三段烃源岩有密切联系;一个埋深超过 4 100 m 的沙三段烃源岩为一类,没有与之密切相关的原油样品;D359-1 原油样品和沙一段烃源岩为一类,油、源间的相关系数为 0.92,表明该油样与沙一段烃源岩密切相关;其余的 DB14-18,D35-5-4,D35-11-x5 和 D355 原油样品为一类,样品



序号	样品名称
1	DX361
2	D359-2
3	Es_4烃源岩
4	Es_4烃源岩
5	Es_4烃源岩
6	DB25-23
7	Es_3烃源岩
8	DB10-4
9	D371
10	DB65-51
11	D65
12	Es_3烃源岩
13	Es_3烃源岩
14	Es_3烃源岩
15	Es_3烃源岩
16	D359-1
17	Es_1烃源岩
18	D35-5-4
19	D35-11-x5
20	D355
21	DB14-18

图 4-9 大王北洼陷甾萜烷油源对比聚类谱系图

间的相关系数超过 0.9，与这类原油样品关系最密切的是沙一段烃源岩。

【例 4】储集层综合评价。

在储集层定量评价中，常用的评价指标有均值、饱和度中值、毛细管压力、渗透率、孔隙度和分选系数等。以上述指标为变量，采用欧氏距离系数为分类统计量，对华北某地震旦系雾迷山组 46 个岩样进行 Q 型聚类分析，得到储集层定量分类谱系图(图 4-10)。

由图 4-10 可以看出，若以距离系数 0.25 为分类标准，特征典型的储集层有三类，即以溶蚀孔洞缝与构造缝为主的好储层、以晶间缝为主的差储集层、以基质微孔为主的非储层，而 1，14，8 号岩样为特殊类型的储集层。

【例 5】岩性段划分。

某盆地 g114 井有自然伽马、自然电位等测井资料。在井段 1 870～2 040 m，对自然伽马、自然电位曲线以 0.25 m 的深度间隔取样构成数据序列，对该数据序列最优分割，绘出岩性剖面与分割结果对比图(图 4-11)。

从图 4-11 上可以看出，基本上可以把岩性段分开，相邻的两条分割线是某种岩性段的顶、底界限。由此看出，对测井曲线进行分解法聚类分析，可以为测井地质解释提供参考依据。

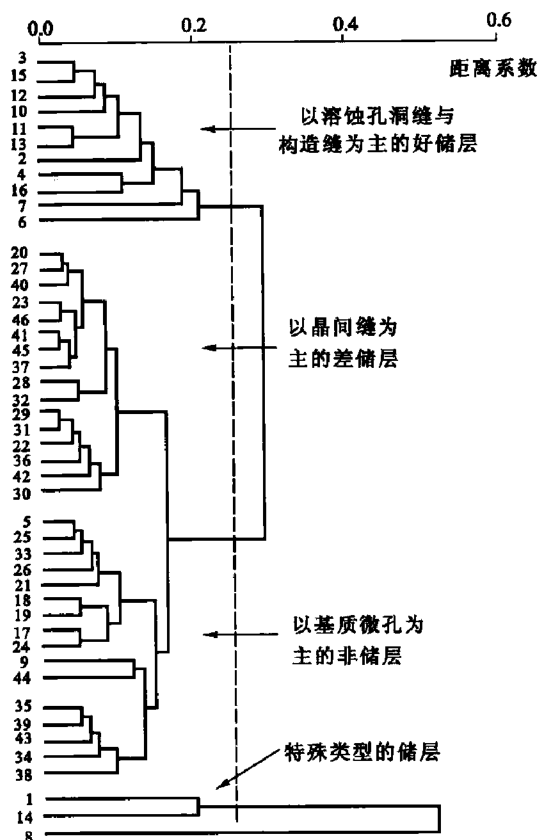


图 4-10 华北某地震旦系雾迷山组储层分类谱系图(据伍友佳,2000,修改)

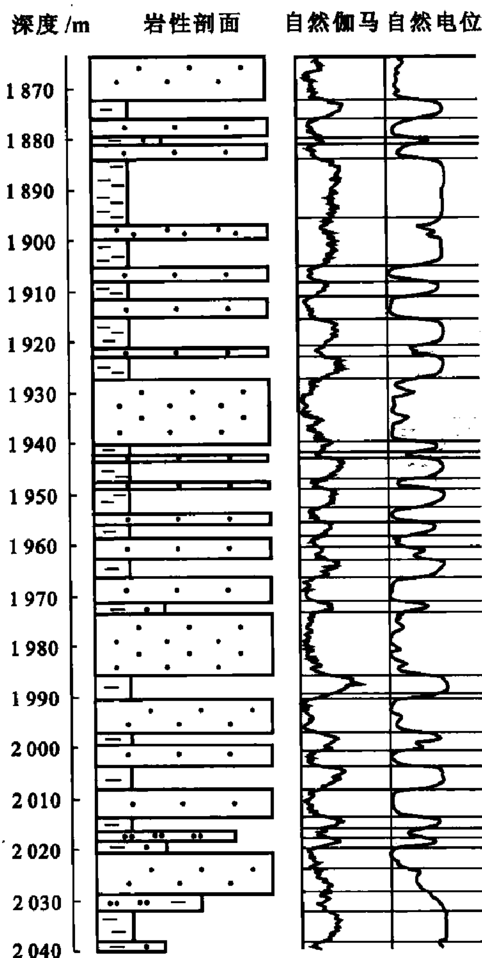


图 4-11 岩性剖面与分割结果对比图

思考与练习

1. 什么是聚类分析?
2. 在聚类分析中为何要对原始数据进行预处理?
3. 常用的定量数据预处理方法有几种?试写出常用的定量数据预处理方法变换公式?
4. 最常用的聚合法聚类统计量是什么?其地质内涵是什么?
5. 分解法聚类分析的统计量是什么?其地质内涵是什么?
6. 试述聚合法与分解法聚类分析的基本过程。
7. 如何确定分解法的最优分段数?
8. 假设 k_1,k_2,k_3 分别是取自三套生油层系的岩样,k_0 是取自储集层的油样。据表 4-2 中的数据,试用相关系数、相似系数、距离系数进行聚合法聚类分析,并分析聚类结果的差异,最终对油样 k_0 的油源做出合理的解释。
9. 现有某井段自然伽马、自然电位、电阻率测井的六组实测值及其正规化数据(表 4-3)。试据表中两种数据对数据序列进行最优 3 分割,并比较分类结果的差异。

表 4-2　甾烷族化合物相对含量表

样品 \ 相对含量 m/z	372	386	398	400	412	424
k_1	3.72	11.20	5.78	15.80	5.20	9.11
k_2	5.30	6.12	5.30	8.10	7.12	7.80
k_3	9.65	10.20	9.61	13.20	9.40	10.70
k_0	10.78	13.10	7.81	15.90	7.81	12.80

表 4-3　某井段电测井数据

测值深度/m	自然伽马/($\times10^{-6}$伦琴·h^{-1})		自然电位/mV		电阻率/(Ω·m)	
	实测值	正规化值	实测值	正规化值	实测值	正规化值
1 885	52.99	0.000 0	2.76	0.710 8	94.85	1.000 0
1 886	63.36	0.005 8	2.83	0.795 2	77.22	0.807 8
1 887	73.36	0.317 0	2.39	0.265 2	60.61	0.626 7
1 888	78.58	0.398 2	2.17	0.000 0	35.77	0.355 9
1 889	77.24	0.377 4	3.00	1.000 0	11.55	0.091 8
1 890	117.25	1.000 0	2.00	0.060 2	3.13	0.000 0

第五章　判别分析

地质学领域内有很多属于归类研究的问题，如钻穿储层的含油气性，岩石样品的沉积相，生油岩的热演化阶段等。这类问题的共性是确定个体应属于已知类中的哪一类，即对个体进行归类，或者说对个体的归属做出判定。为了叙述方便，我们不讨论具体问题，并统称个体和已知的类为样品和总体，在此基础上给出判别分析的一般概念。

假设 $A=\{a_1,a_2,\cdots,a_G\}$ 是从已知的 G 个总体中取出的 G 组样品（代表 G 个总体），每个样品有 m 个变量。判别分析是根据 G 组样品 m 个变量的观测值，建立总体与样品 $X(X\in A)$ 之间的定量关系，即判别函数的一种多元统计分析方法。当 $G=2$ 时，叫做两总体判别，又称为线性判别；当 $G>2$ 时，叫做多总体判别。逐步判别分析是在 m 个变量中"筛选"判别能力强的变量建立判别函数的多元统计分析方法。

§1　两总体判别分析

两总体判别就是确定样品 X 是属于总体 A 还是属于总体 B 的问题。判定样品归属的判别函数叫做线性判别函数。

一、线性判别函数的一般形式

如果样品仅有 x_1,x_2 两个变量，总体 A,B 的样品点分别落在两个椭圆内（图 5-1），当 x_1,x_2 分别落在区间 $(a,b),(c,d)$ 内时，就不能确定样品属于总体 A 还是属于总体 B。如果把坐标系旋转 α 角，变为新的 y,z 坐标系，那么变量 y 可以把总体 A 与 B 分开，即可以用 y 判定样品的归属。变量 y 的形式为：

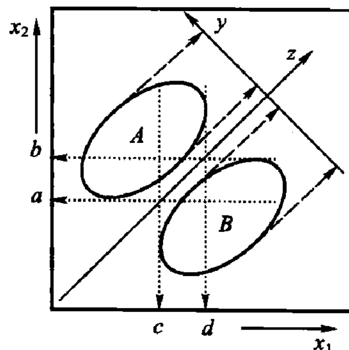


图 5-1　判别分析示意图

$$y=c_1x_1+c_2x_2$$

设样品有 m 个变量，那么 y 的一般形式为：

$$y=c_1x_1+c_2x_2+\cdots+c_mx_m \qquad (5\text{-}1)$$

式(5-1)称为线性判别函数，它是 $m+1$ 维空间的一个平面。

二、确定判别函数的系数

1. *原始数据*

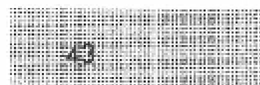
进行线性判别分析的任务之一就是根据样品观测值确定式(5-1)中的系数 $c_1,c_2,\cdots,c_m$。假设从总体 A,B 中分别取出 n_a,n_b 个样品，每个样品有 m 个变量，它们的观测值分别记为：

$$x_{ij}(a),x_{kj}(b)(i=1,2,\cdots,n_a;k=1,2,\cdots,n_b;j=1,2,\cdots,m) \qquad (5\text{-}2)$$

这是建立线性判别函数的原始数据。

2. *费歇尔准则下的判别函数*

把式(5-2)中的观测值分别代入式(5-1)，得判别函数值：



$$y_i(a) = \sum_{j=1}^{m} c_j x_{ij}(a) \quad (i = 1,2,\cdots,n_a)$$

$$y_k(b) = \sum_{j=1}^{m} c_j x_{kj}(b) \quad (k = 1,2,\cdots,n_b)$$

记

$$Q = [\bar{y}(a) - \bar{y}(b)]^2 \tag{5-3}$$

$$H = \sum_{i=1}^{n_a}[y_i(a) - \bar{y}(a)]^2 + \sum_{k=1}^{n_b}[y_k(b) - \bar{y}(b)]^2 \tag{5-4}$$

式(5-3),(5-4)中

$$\bar{y}(a) = \frac{1}{n_a}\sum_{i=1}^{n_a} y_i(a) = \sum_{j=1}^{m} c_j \bar{x}_j(a)$$

$$\bar{y}(b) = \frac{1}{n_b}\sum_{k=1}^{n_b} y_k(b) = \sum_{j=1}^{m} c_j \bar{x}_j(b)$$

建立判别函数时要求 Q 达到最大，H 达到最小(图5-2)，即两组判别函数点的中心距最大，组内判别函数点的离散度最小。满足以上条件的判别函数可最大限度地把 A,B 区分开。上述准则由费歇尔提出，故称费歇尔准则。

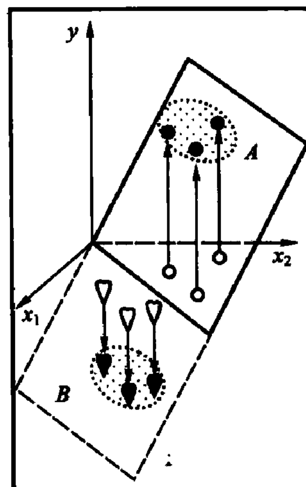


图 5-2 样品点在 y 平面上的投影

上述准则等价于要求

$$V = Q/H$$

达到最大。V 是 $c_j(j=1,2,\cdots,m)$ 的二次函数，且 $V>0$，根据极值原理有：

$$\frac{\partial V}{\partial c_j} = 0 \quad (j = 1,2,\cdots,m)$$

对上式化简整理，则有：

$$\sum_{k=1}^{m} s_{jk} c_k = d_j \tag{5-5}$$

式中

$$s_{jk} = \sum_{i=1}^{n_a}[x_{ij}(a) - \bar{x}_j(a)][x_{ik}(a) - \bar{x}_k(a)] + \sum_{i=1}^{n_b}[x_{ij}(b) - \bar{x}_j(b)][x_{ik}(b) - \bar{x}_k(b)] \quad (j,k = 1,2,\cdots,m)$$

$$d_j = [\bar{x}_j(a) - \bar{x}_j(b)] \quad (j = 1,2,\cdots,m)$$

式(5-5)是以 c_j 为变量的方程组，从中可解出 c_j 得判别函数式(5-1)。

三、显著性检验及样品的判别

若总体 A,B 差异不明显，那么由观测值建立的判别函数就没有实际意义。为此，需要对 A,B 的差异性进行检验。

1. 检验方法

用已建立的判别函数对已知样品的总体重新判定，若判断对了 $n(n\leqslant(n_a+n_b))$ 个，定

义 $r=n/(n_a+n_b)$为对判率。r 越大，A,B 差异就越明显，判别函数的判别效果就越显著。

2. 样品总体的判别

在判别函数显著的条件下，定义

$$y_c=[n_a\bar{y}(a)+n_b\bar{y}(b)]/(n_a+n_b)$$

为判别样品总体的判别指数。

若 $\bar{y}(a)>y>\bar{y}(b)$，把样品 X 的观测值 $x_j(j=1,2,\cdots,m)$代入判别函数式(5-1)，得判别函数值 y，当 $y>y_c$ 时，$X\in A$，否则 $X\in B$。

§2 多总体判别分析

一、原始数据

如果从 G 个总体中分别取出 $n_g(g=1,2,\cdots,G)$个样品，每个样品有 m 个变量，那么样品观测值构成的观测样本为：

$$\boldsymbol{X}_{gk}=\begin{pmatrix}x_{gk}^{(1)}\\x_{gk}^{(2)}\\\vdots\\x_{gk}^{(m)}\end{pmatrix}\qquad(g=1,2,\cdots,G;k=1,2,\cdots,n_g)$$

式中 $x_{gk}^{(i)}$——总体 $a_g(g=1,2,\cdots,G)$中第 $k(k=1,2,\cdots,n_g)$个样品的第 i 个变量的观测值。

$\boldsymbol{X}_{gk}$——求判别函数的原始数据。

二、贝叶斯(Bayes)准则下建立多总体判别函数的基本原理

对于取自 G 个已知总体的样品 X，在对它所属的总体作出判定以前，它可能属于任何一个总体，但是它归属于总体 $a_g(g=1,2,\cdots,G)$的概率不同。由 Bayes 公式可以求得 X 属于总体 $a_g(g=1,2,\cdots,G)$的条件概率为：

$$\begin{aligned}p(a_g/X)&=p(a_g)p(X/a_g)\Big/\sum_{j=1}^{G}p(a_j)p(X/a_j)\\&=p_gf_g(X)\Big/\sum_{j=1}^{G}p_jf_j(X)\end{aligned}\qquad(5\text{-}6)$$

式中 $p_g,f_g(X)$——总体 $a_g(g=1,2,\cdots,G)$的先验概率和概率密度。

如果 $p(a_k/X)$是条件概率中的最大者，即

$$p(a_k/X)=\max_{i\leqslant g\leqslant G}p(a_g/X)$$

那么，判定样品 $X\in a_k$ 时判错的概率最小。在计算条件概率 $p(a_g/X)$时，式(5-6)的分母是一个常数，条件概率的相对大小不会受其影响。因此，记

$$E_g(X)=p_gf_g(X)\qquad(g=1,2,\cdots,G)\qquad(5\text{-}7)$$

式(5-7)是 Bayes 准则下多总体判别的一般判别函数，根据函数值的相对大小可对样品 X 的总体做出判别。

三、正态总体的判别函数

用式(5-7)判定样品 X 的总体，需要进一步确定总体的先验概率 p_g 和概率密度 $f_g(X)$。

假设总体服从正态分布，其概率密度为：

$$f_g(\boldsymbol{X}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_g)\right] \tag{5-8}$$

$$\boldsymbol{X} = (x^{(1)}, x^{(2)}, \cdots, x^{(m)})'$$

式中 $\boldsymbol{\mu}_g$——a_g 的期望向量；

$\boldsymbol{\Sigma}$——各个总体共同的协方差矩阵；

$\boldsymbol{\Sigma}^{-1}$——$\boldsymbol{\Sigma}$ 的逆矩阵。

由原始数据可求得 $\boldsymbol{\mu}_g$，$\boldsymbol{\Sigma}$ 的估计值 $\overline{\boldsymbol{X}}_g$ 和 $\boldsymbol{S}$，并且

$$\overline{\boldsymbol{X}}_g = \begin{pmatrix} \bar{x}_g^{(1)} \\ \bar{x}_g^{(2)} \\ \vdots \\ \bar{x}_g^{(m)} \end{pmatrix} \quad (g = 1,2,\cdots,G); \qquad \boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}$$

式中

$$\bar{x}_g^{(i)} = \frac{1}{n_g}\sum_{k=1}^{n_g} x_{gk}^{(i)} \qquad (i = 1,2,\cdots,m)$$

$$s_{ij} = \frac{1}{N-G}\sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)} - \bar{x}_g^{(i)})(x_{gk}^{(j)} - \bar{x}_g^{(j)})(i,j = 1,2,\cdots,m; N = n_1 + n_2 + \cdots + n_G)$$

由此，可把式(5-8)改写为：

$$f_g(\boldsymbol{X}) = \frac{|\boldsymbol{S}^{-1}|^{1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(\boldsymbol{X}-\overline{\boldsymbol{X}}_g)'\boldsymbol{S}^{-1}(\boldsymbol{X}-\overline{\boldsymbol{X}}_g)\right] \tag{5-9}$$

把式(5-9)和 $p_g = n_g/N$ 代入式(5-7)，再对该式两边取自然对数并舍去其中与 g 无关的项，化简得正态总体下的判别函数：

$$\begin{aligned} F_g(\boldsymbol{X}) &= \ln q_g + \boldsymbol{X}'\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}_g - \frac{1}{2}\overline{\boldsymbol{X}}_g'\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}_g \\ &= \ln q_g + \sum_{k=1}^{m} c_g^{(k)} x^{(k)} + c_{0g} \qquad (g = 1,2,\cdots,G) \end{aligned} \tag{5-10}$$

式中

$$c_g^{(k)} = \sum_{l=1}^{m} s_{kl}^{-1}\bar{x}_g^{(l)}$$

$$c_{0g} = -\frac{1}{2}\sum_{l=1}^{m} c_g^{(k)} \cdot \bar{x}_g^{(k)} \qquad (k = 1,2,\cdots,m)$$

对于服从其他分布的总体来说，仿照上述做法可以得到相应的判别函数。

四、对样品总体的判别

把样品 X 的观测值 $\boldsymbol{X} = (x^{(1)}, x^{(2)}, \cdots, x^{(m)})'$ 代入式(5-10)，得 $F_g(\boldsymbol{X})(g=1,2,\cdots,G)$，若

$$F_k(\boldsymbol{X}) = \max_{1 \leqslant g \leqslant G} F_g(\boldsymbol{X})$$

那么判定样品 X 属于总体 a_k 的条件概率为：

$$p_k = \exp[F_k(\boldsymbol{X})] \Big/ \sum_{j=1}^{G}\exp[F_j(\boldsymbol{X})] \qquad (k = 1,2,\cdots,G)$$

五、判别函数的显著性检验

1. *对判率检验*

利用式(5-10)对观测样本中 $N(N=n_1+n_2+\cdots+n_G)$ 个样品的总体重新判定，若判断对了 $n(n\leqslant N)$ 个，那么称 $r(r=n/N)$ 为对判率。r 越大，总体间的差异就越明显，判别函数的判别效果就越好。

2. *马哈拉诺比斯距离 D^2 检验*

假设 H_0：总体差异不明显。

统计量为：

$$D^2=\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{G}n_k s_{ij}^{-1}(\bar{x}_k^{(i)}-\bar{x}^{(i)})(\bar{x}_k^{(j)}-\bar{x}^{(j)})$$

式中

$$\bar{x}^{(i)}=\frac{1}{N}\sum_{k=1}^{G}\sum_{j=1}^{n_k}x_{kj}^{(i)}=\frac{1}{N}\sum_{k=1}^{G}n_k\cdot\bar{x}_k^{(i)}\qquad(i=1,2,\cdots,m)$$

D^2 服从自由度为 $m(G-1)$ 的 χ^2 分布。给定检验水平 α，查 χ_α^2 分布表得 D^2 的临界值 D^*。当 $D^2>D^*$ 时，否定假设 H_0，即拟定的 m 个变量能够区分开已知的 G 个总体，否则接受假设 H_0，即拟定的 m 个变量不能对样品的归属做出正确的判别，此时应剔除其中区分能力小的变量或者引入一些更有效的变量，重新建立判别函数。

§3 逐步判别分析

在拟定的判别变量 $x^{(1)},x^{(2)},\cdots,x^{(m)}$ 之间既有相对的独立性，又存在着一定的成因联系。对于区分已知总体来说，具有成因联系的那些变量表面上看其各自的区分能力都不可忽视，但当把 $x^{(i)}$ 选入判别函数后，又使得先选入的 $x^{(a)}$ 的区分能力变得不显著了。另外，建立判别函数时需要求 $\boldsymbol{S}^{-1}$，若存在区分能力不显著的变量，将导致 $\boldsymbol{S}^{-1}$ 不存在，故求不出判别函数。鉴于上述原因，提出“筛选”判别能力强的变量，建立经济实用的判别函数的逐步判别分析。它的基本思想与逐步回归分析类似，即逐个检验变量的区分能力，把区分能力强的变量“引入”判别函数，在引入变量的过程中，随时“剔除”已引入判别函数中的区分能力变弱的变量，直到既没有区分能力强的变量引入，又没有区分能力变弱的变量剔除为止。

一、逐步判别分析方法原理

1. *原始数据*

逐步判别分析的原始数据见本章§2中的原始数据。

2. *Wilks Λ 统计量*

它是检验变量区分能力的指标。假设样本数据来自 G 个具有相同协方差矩阵的正态总体 $N(\mu_g,\boldsymbol{\Sigma})$。为了检验变量的区分能力，定义样本内离差矩阵 $\boldsymbol{W}$、样本间离差矩阵 $\boldsymbol{B}$、总离差矩阵 $\boldsymbol{T}$：

样本 a_g 中第 i 个变量的平均值 $\bar{x}_g^{(i)}=\dfrac{1}{n_g}\sum\limits_{k=1}^{n_g}x_{gk}^{(i)}\ (g=1,2,\cdots,G)$

全部样品第 i 个变量的平均值 $\bar{x}^{(i)}=\sum\limits_{g=1}^{G}\sum\limits_{k=1}^{n_g}x_{gk}^{(i)}\Big/\sum\limits_{g=1}^{G}n_g=\sum\limits_{g=1}^{G}n_g\bar{x}_g^{(i)}\Big/\sum\limits_{g=1}^{G}n_g$

$$w_{ij}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)}-\bar{x}_g^{(i)})(x_{gk}^{(j)}-\bar{x}_g^{(j)})$$

$$b_{ij}=\sum_{g=1}^{G}n_g(\bar{x}_g^{(i)}-\bar{x}^{(i)})(\bar{x}_g^{(j)}-\bar{x}^{(j)})$$

$$t_{ij}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)}-\bar{x}^{(i)})(x_{gk}^{(j)}-\bar{x}^{(j)})$$

$$\boldsymbol{W}=(w_{ij})_{m\times m},\qquad \boldsymbol{B}=(b_{ij})_{m\times m},\qquad \boldsymbol{T}=(t_{ij})_{m\times m}$$

可以证明：

$$\boldsymbol{T}=\boldsymbol{W}+\boldsymbol{B}$$

Wilks Λ 统计量

$$U=|\boldsymbol{W}|/|\boldsymbol{T}|$$

是在假设

$$H_0:\mu_1=\mu_2=\cdots=\mu_G$$

下检验全部变量综合区分能力的统计指标。U 越小，样本内部差异越小，样本之间差异越大，即 H_0 不成立。

3. “引入”与“剔除”变量的统计量

统计量 U 是两个行列式的比，如果按行列式列号 $r_1,r_2,\cdots,r_m$ 的顺序对行列式进行消去计算，并表示出消去次序，那么 U 可以改写为：

$$U_{r_1r_2\cdots r_m}=[w_{r_1r_1}^{(0)}w_{r_2r_2}^{(1)}\cdots w_{r_mr_m}^{(m-1)}][t_{r_1r_1}^{(0)}t_{r_2r_2}^{(1)}\cdots t_{r_mr_m}^{(m-1)}]^{-1}\tag{5-11}$$

在式(5-11)的基础上，可以导出检验变量 $x^{(r)}$ 判别能力的 Wilks Λ 统计量。

(1) “引入”变量 $x^{(r)}$ 的 Wilks Λ 统计量及对 $x^{(r)}$ 的检验。

假设判别分析进行了 p 步，共引入了 p 个变量 $x^{(r_1)},x^{(r_2)},\cdots,x^{(r_p)}$（都是判别能力强的变量），根据式(5-11)，有：

$$U_{r_1r_2\cdots r_p}=[w_{r_1r_1}^{(0)}w_{r_2r_2}^{(1)}\cdots w_{r_pr_p}^{(p-1)}][t_{r_1r_1}^{(0)}t_{r_2r_2}^{(1)}\cdots t_{r_pr_p}^{(p-1)}]^{-1}\tag{5-12}$$

若判别分析的第 $p+1$ 步是在判别函数中再引入变量 $x^{(r)}$，则有：

$$U_{r_1r_2\cdots r_pr}=[w_{r_1r_1}^{(0)}w_{r_2r_2}^{(1)}\cdots w_{r_pr_p}^{(p-1)}w_{rr}^{(p)}][t_{r_1r_1}^{(0)}t_{r_2r_2}^{(1)}\cdots t_{r_pr_p}^{(p-1)}t_{rr}^{(p)}]^{-1}\tag{5-13}$$

由式(5-12)，(5-13)可知，$w_{rr}^{(p)}/t_{rr}^{(p)}$ 是引入变量 $x^{(r)}$ 后 U 的改变因子，将其记为：

$$U_r=w_{rr}^{(p)}/t_{rr}^{(p)}\qquad (r\notin(r_1,r_2,\cdots,r_p))\tag{5-14}$$

U_r 越小，表明变量 $x^{(r)}$ 在样本之间的差异越明显，即它的判别能力就越强。因此，U_r 是检验变量 $x^{(r)}$ 判别能力的 Wilks Λ 统计量。

假设变量 $x^{(r)}$ 的判别能力小，则统计量

$$\begin{aligned}F_1&=[(1-U_r)/(G-1)]/[U_r/(N-G-p)]\\&=[(N-G-p)(t_{rr}^{(p)}-w_{rr}^{(p)})]/[(g-1)w_{rr}^{(p)}]\end{aligned}$$

服从第一自由度为$(G-1)$、第二自由度为$(N-G-p)$的 F 分布。对于给定检验水平 α，查分布表 $F_\alpha(G-1,N-g-p)$得临界值 F_α，若 $F_1>F_\alpha$，则变量 $x^{(r)}$ 的判别能力强，应把变量 $x^{(r)}$ 引入判别函数。

(2) “剔除”变量 $x^{(r)}$ 的 Wilks Λ 统计量及对 $x^{(r)}$ 的检验。

假设判别分析进行了 p 步，共引入了 p 个变量 $x^{(r_1)},x^{(r_2)},\cdots,x^{(r_p)}$（没有被剔除的变量），若判别分析的第 $p+1$ 步是拟剔除判别函数中的变量 $x^{(r)}$($r\in(r_1,r_2,\cdots,r_p)$)，那么 $x^{(r)}$

的判别能力可视为第 p 步引入 $x^{(r)}$ 的判别能力，即：

$$U_r^* = w_{rr}^{(p-1)}/t_{rr}^{(p-1)} \qquad (r \in (r_1, r_2, \cdots, r_p)) \tag{5-15}$$

假设变量 $x^{(r)}$ 的判别能力小，则统计量

$$\begin{aligned} F_2 &= [(1-U_r^*)/(G-1)]/[U_r^*/(N-G-p+1)] \\ &= [(N-G-p+1)(1-U_r^*)]/[(G-1)U_r^{(*)}] \end{aligned}$$

服从第一自由度为 $(G-1)$、第二自由度为 $(N-G-p+1)$ 的 F 分布。给定检验水平 α，查 $F_\alpha(G-1, N-G-p+1)$ 分布表，得临界值 F_α^*，若 $F_2 \leqslant F_\alpha^*$，则变量 $x^{(r)}$ 的判别能力小，应从判别函数中剔除变量 $x^{(r)}$。

4. 逐步判别分析的变换公式

逐步判别分析求解判别函数的过程与逐步回归分析求解回归方程的过程类似，不同之处是逐步回归分析只是对相关系数增广矩阵 $\boldsymbol{R}$ 进行变换，而逐步判别分析要对 $\boldsymbol{W}$ 和 $\boldsymbol{T}$ 两个矩阵进行变换。逐步判别分析的第 $p+1$ 步不论是引入还是剔除变量 $x^{(r)}$，都是按式(5-16)，(5-17)对 $\boldsymbol{W}, \boldsymbol{T}$ 矩阵进行一次变换。第 $p+1$ 步变换消去 $\boldsymbol{W}, \boldsymbol{T}$ 矩阵中第 r 列的变换公式如下：

$$w_{kl}^{(p+1)} = \begin{cases} 1/w_{rr}^{(p)} & (k=r, l=r) \\ w_{kl}^{(p)}/w_{rr}^{(p)} & (k=r, l\neq r) \\ -w_{kr}^{(p)}/w_{rr}^{(p)} & (k\neq r, l=r) \\ w_{kl}^{(p)} - w_{kr}^{(p)} \cdot w_{rl}^{(p)}/w_{rr}^{(p)} & (k\neq r, l\neq r) \end{cases} \tag{5-16}$$

$$t_{kl}^{(p+1)} = \begin{cases} 1/t_{rr}^{(p)} & (k=r, l=r) \\ t_{kl}^{(p)}/t_{rr}^{(p)} & (k=r, l\neq r) \\ -t_{kr}^{(p)}/t_{rr}^{(p)} & (k\neq r, l=r) \\ t_{kl}^{(p)} - t_{kr}^{(p)} \cdot t_{rl}^{(p)}/t_{rr}^{(p)} & (k\neq r, l\neq r) \end{cases} \tag{5-17}$$

二、判别函数和对样品的判别

1. 判别函数

若逐步判别分析进行了 p 步结束，引入变量数为 $v(v\leqslant m)$，那么判别函数为：

$$F_g(\boldsymbol{X}) = \ln q_g + \sum_{i\in v} c_g^{(i)} x^{(i)} + c_{0g} \tag{5-18}$$

式中

$$\begin{cases} c_g^{(i)} = (N-G)\sum_{j\in v} w_{ij}^{(p)} \bar{x}_g^{(j)} & (i\in v, g=1,2,\cdots,G) \\ c_{0g} = -\dfrac{1}{2}\sum_{i\in v} c_g^{(i)} \bar{x}_g^{(i)} & (g=1,2,\cdots,G) \end{cases}$$

2. 对样品的判别

把 $\boldsymbol{X}=(x^{(1)}, x^{(2)}, \cdots, x^{(m)})'$ 代入式(5-18)，得判别函数值 $F_g(\boldsymbol{X})(g=1,2,\cdots,G)$，若

$$F_k(\boldsymbol{X}) = \max_{1\leqslant g\leqslant G} F_g(\boldsymbol{X})$$

那么样品 $X \in a_k$。$X \in a_k$ 的条件概率为：

$$p_k = \exp[F_k(\boldsymbol{X})]/\sum_{j=1}^{G}\exp[F_j(\boldsymbol{X})]$$

三、逐步判别分析计算过程

总结逐步判别分析的计算过程，给出逐步判别分析流程图(图 5-3)。

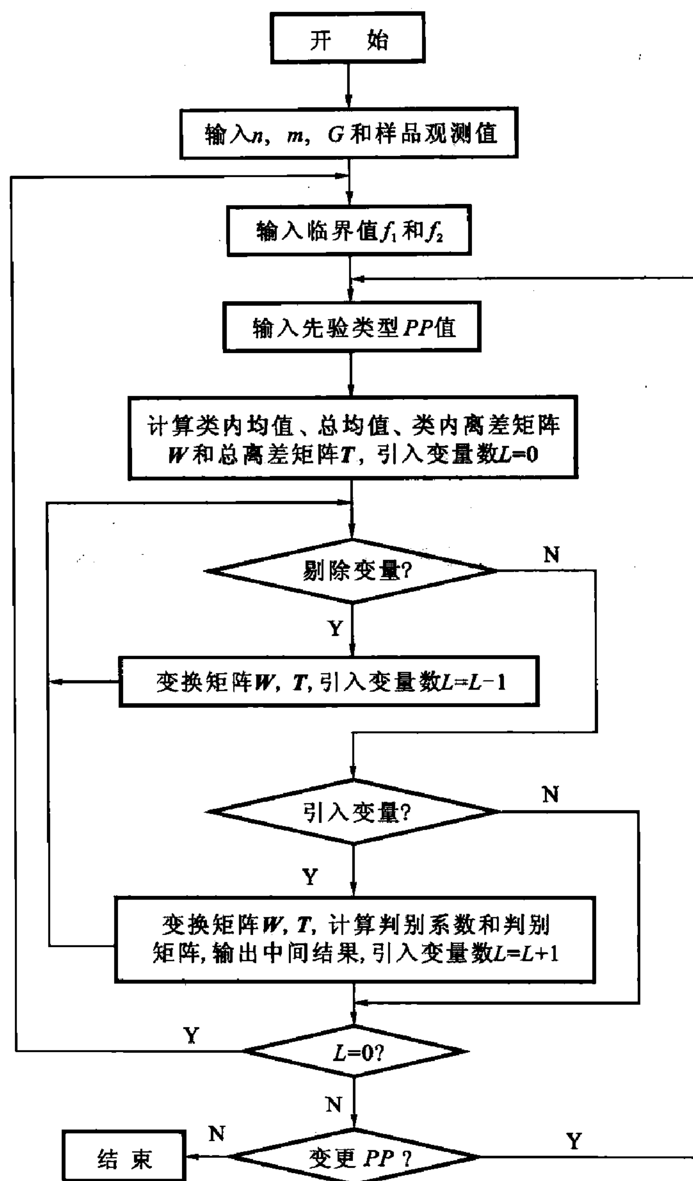


图 5-3　判别分析流程图

§4　应用实例

【例 1】岩性剖面的反演。

测井参数是岩石的效应，其观测值的差异主要取决于岩性，即依赖于岩石的矿物成分，颗粒的大小、结构以及岩石孔隙中流体的性质。也就是说，根据钻孔的测井参数，可以反演钻孔的岩性剖面。

某砂砾岩油田，岩心实物少，但测井资料比较丰富，多数井均有如下测井资料：

x_1——微电极电阻率，Ω·m；

x_2——2.5 m 梯度电阻率，Ω·m；

x_3——4 m 梯度电阻率，Ω·m；

x_4——感应电导率，mS/m；

x_5——声波时差，$\mu s/m$；

x_6——浅侧向，$\Omega \cdot m$；

x_7——补偿中子孔隙度，%；

x_8——井径，cm；

x_9——微电极差，$\Omega \cdot m$。

为了开展砂砾岩油藏描述工作，利用上述测井资料建立了岩性识别函数，反演了30余口无岩心井的岩性剖面，为沉积相和储层研究提供了资料。具体做法如下：

(1) 分析砂砾岩油田的岩心，确定岩石类型数目，在有岩心的钻孔剖面上采集不同岩性对应的各种测井参数值，作为判别分析的样本。

(2) 根据不同岩石类型的样本值，建立识别岩性的判别函数。

(3) 把具有岩心井段的测井参数曲线离散抽样输入计算机，利用已建立的判别函数对有岩心井段的岩性进行识别，以检验判别函数的有效性。

(4) 在判别函数有效的条件下，把无岩心井的相应测井参数曲线离散抽样输入计算机。

(5) 把不同深度点上各测井参数的离散抽样值代入判别函数，计算判别函数值，以其中最大者对采样点的岩性进行归类，并记录下归类号、相应的深度和测井参数。

(6) 根据上一步记录的资料，由计算机绘制岩性剖面、相应的测井参数曲线以及所需的等值线图等。

某砂砾岩油田的岩心可分为砾岩、砂岩和泥岩三种类型。在岩性剖面上取了可靠的84个样品（其中砾岩样品30个，砂岩样品30个，泥岩样品24个）及其对应的9项测井参数，进行逐步判别分析，引入7项参数，得到识别砾岩、砂岩和泥岩的判别函数如下：

$$\begin{aligned}F_1(\boldsymbol{X}) = &\ 1.785\,6x_1 + 0.646\,5x_3 + 0.155\,8x_4 + 23.603\,6x_6 \\ &+ 1.756\,1x_7 + 14.506\,0x_8 - 0.202\,7x_9 - 205.392\,0\end{aligned} \quad \text{（砾岩判别函数）}$$

$$\begin{aligned}F_2(\boldsymbol{X}) = &\ 1.126\,9x_1 + 0.479\,4x_3 + 0.150\,6x_4 + 16.749\,6x_6 \\ &+ 1.873\,2x_7 + 14.869\,5x_8 - 2.429\,9x_9 - 197.860\,5\end{aligned} \quad \text{（砂岩判别函数）}$$

$$\begin{aligned}F_3(\boldsymbol{X}) = &-0.054\,5x_1 + 0.313\,5x_3 + 0.203\,2x_4 + 18.949\,7x_6 \\ &+ 2.615\,7x_7 + 17.857\,8x_8 - 4.542\,7x_9 - 287.994\,0\end{aligned} \quad \text{（泥岩判别函数）}$$

上述判别函数对84个岩石样品中砾岩、砂岩、泥岩的对判率分别为93%，97%和96%，平均对判率为95%。对判率从一个方面反映了判别函数识别岩性的可靠程度。但是，这些岩石样品取自不同井的岩心，它们并不代表一段连续的地层剖面。因此，对判率还不能充分说明判别函数对连续地层剖面的识别效果。为此，又用上述判别函数识别了研究区内永1-5井2 226～2 276 m岩心段的岩性，以此来检验判别函数对连续地层剖面进行岩性识别的有效性。

从岩心剖面与识别的地层岩性剖面（预测剖面，图5-4）上看出，两个剖面在局部上存在着岩性的差异，造成这一差异的主要原因是利用砾岩、砂岩和泥岩的代表性样品所确定的判别函数对过渡性岩层进行识别时，就会出现岩性上的偏差。另外，上、下岩层对簿夹层测井参数的影响以及测井参数采样间隔偏大也是造成岩性误识的因素。从整体上看，两个剖面吻合较好，预测剖面基本上反映了地层的岩性特征。在上述分析的基础上，利用建立的判别函数预测了研究区内无岩心井的岩性剖面。

【例2】预测生油岩热演化阶段。

深度/m | 岩心剖面 | 预测剖面 | 微电极电阻率/(Ω·m) | 4 m梯度电阻率/(Ω·m) | 感应电导率/(mS·m^{-1}) | 浅侧向/(Ω·m) | 补偿中子孔隙度/% | 井径/cm | 微电极差/(Ω·m)

0.3 5.7 5.8 53.0 6.8 54.0 18.0 235.0 1.8 39.5 0.2 5.0 13.5 32.0

2 226
2 230
2 234
2 238
2 242
2 246
2 250
2 254
2 258
2 262
2 266
2 270
2 274

图 5-4 岩性剖面及部分电测曲线图

根据生油岩的成熟度，一般将生油岩的热演化过程分为未成熟、成熟、高成熟、过成熟四个阶段。确定生油岩的热演化阶段是油气资源评价中的重要研究内容之一，为了定量确定生油岩的热演化阶段，统计了我国有关探区 66 个生油层系的生油门限时间 t、生油层温度 T 以及生油层的埋藏深度 H 的数据(表 5-1)。

根据统计资料，取 $T+273, t, H, 1/H, \ln(T+273), 1/(T+273)$ 为变量，取引入和剔除变量的临界值 $f_1=f_2=0$，引入 4 个变量，得生油岩热演化阶段判别函数：

$$F_1(\boldsymbol{X})=-431.677\,2x_1+4.399\,0x_2-0.261\,0x_3+200\,298.2x_5-510\,438.9 \quad (未成熟)$$

$$F_2(\boldsymbol{X})=-432.683\,6x_1+4.404\,3x_2-0.261\,0x_3+200\,782.5x_5-512\,924.1 \quad (成熟)$$

$$F_3(\boldsymbol{X})=-433.834\,0x_1+4.410\,8x_2-0.260\,8x_3+201\,345.2x_5-515\,827.7 \quad (高成熟)$$

$$F_4(\boldsymbol{X})=-434.446\,5x_1+4.414\,0x_2-0.259\,8x_3+201\,681.6x_5-517\,606.3 \quad (过成熟)$$

表 5-1 生油岩热演化参数

演化阶段	样品序号	地 区	热演化参数		
			(T+273)/K	t/Ma	H/m
未成熟阶段	1	松辽盆地(青 2+3)	337	125.005	1 000
	2	松辽盆地(青 1)	328	123	1 000
	3	松辽盆地(青 2)	334	125	1 750
	4	岐口凹陷	347	33.75	1 680
	5	泌阳凹陷	342	27.999	1 460
	6	湖北(二叠系)	338	242	3 200
	7	潜江凹陷	343	34.38	1 600
	8	高邮凹陷	348	15	1 980
	9	惠民凹陷	351	8.04	1 350
	10	沾化凹陷	344	10.76	1 600
	11	东明凹陷	352	8.5	2 250
	12	松辽盆地(姚 2)	228.5	127	1 480
成熟阶段	13	松辽盆地(青 2+3)	341	127	1 550
	14	松辽盆地(青 2+1)	367	127	2 190
	15	松辽盆地(青)	362	124.2	1 799
	16	松辽盆地(青 2)	364	120.4	1 851
	17	松辽盆地(姚 2)	370	120.4	1 970
	18	岐口凹陷	359	35	2 001
	19	泌阳凹陷	353	31	1 700
	20	泌阳凹陷	367	31	2 099
	21	辽河凹陷	364	48.5	2 001
	22	东台凹陷(阜宁组)	364	34	3 201
	23	湖北(二叠系)	348	257.6	3 450
	24	高邮凹陷(集宁组)	356	17.5	2 200
	25	高邮凹陷(集宁组)	360	18	2 700
	26	沾化凹陷	367	10.7	2 300
	27	沾化凹陷	364	12.8	2 200
	28	沾化凹陷	368	13	2 500
	29	东明凹陷	362	11.6	3 500
高成熟阶段	30	松辽盆地(青 2+3)	380	129	3 100
	31	松辽盆地(青 1)	377	127.88	2 350
	32	松辽盆地(青 1)	402	127.88	2 499
	33	松辽盆地(姚 2)	383	124.64	2 299
	34	松辽盆地(姚 2)	377	121.80	2 150
	35	松辽盆地(姚 2)	389	125.008	2 401
	36	岐口凹陷	419	36	3 701
	37	湖北(二叠系)	392	271.25	4 050
	38	泌阳凹陷	376	34.01	2 300
	39	泌阳凹陷	392	34.05	2 799
	40	东台凹陷(阜宁组)	410	54.25	3 201
	41	东台凹陷(阜宁组)	398	37.005	2 900
	42	东台凹陷(阜宁组)	403	35.5	3 701
	43	湖北(二叠系)	399	271.25	4 001
	44	高邮凹陷(集宁组)	380	23.5	3 000
	45	高邮凹陷(集宁组)	382	24	3 400
	46	惠民凹陷	383	13	2 800
	47	惠民凹陷	386	13.5	3 100
	48	高邮凹陷(集宁组)	399	25	4 100
	49	沾化凹陷	381	16	2 700
	50	沾化凹陷	386	16.5	3 000
	51	沾化凹陷	387	16.5	3 100
	52	东明凹陷	410	15.05	3 550
	53	东明凹陷	712	18	3 550
	54	松辽盆地(青 2+3)	400	130	3 400

续表

演化阶段	样品序号	地　区	热演化参数		
			(T+273)/K	t/Ma	H/m
过成熟阶段	55	松辽盆地(姚 2)	434	123.205	3 555
	56	江汉盆地(潜江组)	440	37.005	4 300
	57	东台凹陷(阜宁组)	444	60.005	4 100
	58	东台凹陷(阜宁组)	433	57.005	5 100
	59	湖北(二叠系)	469	285	5 701
	60	湖北(二叠系)	572	285.5	7 199
	61	高邮凹陷(集宁组)	403	26	4 001
	62	东明凹陷	435	20	3 700
	63	东明凹陷	422	18	3 800
	64	泌阳凹陷	433	36.05	4 140
	65	湖北(二叠系)	454	271.29	5 200
	66	湖北(二叠系)	440	271.29	4 900

变量在判别函数中的引入顺序以及各演化阶段的对判率见表 5-2。

由变量的引入顺序看出,温度是有机质热演化过程的决定因素,温度不足可以在临界温度下通过演化时间来补偿。各演化阶段的对判率都超过了 80%,故可用上述判别函数判别生油岩的热演化阶段。

表 5-2　变量引入顺序及演化阶段正判率

引入顺序	变量代号	变量名	F 检验量	演化阶段	正判率/%
1	x_5	ln(T+273)	90.107 8	未成熟	83
2	x_1	T+273	29.072 5	成　熟	94
3	x_2	t	0.433 9	高成熟	96
4	x_3	H	0.297 9	过成熟	92

珠江口盆地第三系生油岩为中新世至晚渐新世,地层的绝对年龄约 16～30 Ma,埋藏深度为 2 200 m,地层温度为 104 ℃。若地层绝对年龄以 25 Ma 计,按上述判别函数计算,则有:$F_1(\boldsymbol{X}) = 514\ 572.6$,$F_2(\boldsymbol{X}) = 514\ 581.6$,$F_3(\boldsymbol{X}) = 514\ 582.5$,$F_4(\boldsymbol{X}) = 514\ 570.8$,其中最大值为 $F_3(\boldsymbol{X})$。因此,可以认为珠江口盆地第三系生油岩处在高成熟阶段。

【例 3】识别沉积相。

不同沉积环境下形成的沉积物,其成分成熟度和粒度参数不同,因此,可以根据沉积岩样品的成分成熟度或者粒度参数,反推样品的沉积相。如东濮凹陷西部沙三段有三角洲、浊流和风暴流三种沉积相。表 5-3 统计了上述沉积相中 45 块岩样的成分成熟度参数 x_1(石英/(长石+岩屑))、杂基含量 x_2 和胶结物含量 x_3,同时还分析了样品的粒度参数 M_z,σ,S_{k1} 和 k_g。根据表 5-3 中的数据,分别建立了三角洲、浊流和风暴流相识别函数。

成分成熟度参数判别函数为:

$$F_1(\boldsymbol{X}) = 2.284\ 1x_1 + 35.360\ 9x_2 + 40.639\ 0x_3 - 10.305\ 0 \quad \text{(三角洲相)}$$

$$F_2(\boldsymbol{X}) = 1.132\ 5x_1 + 45.340\ 7x_2 + 25.947\ 9x_3 - 7.493\ 4 \quad \text{(浊流相)}$$

$$F_3(\boldsymbol{X}) = 1.606\ 0x_1 + 33.276\ 2x_2 + 36.221\ 3x_3 - 7.964\ 5 \quad \text{(风暴流相)}$$

粒度参数判别函数为:

表 5-3 沉积岩样品参数及判别检验结果

沉积相	样品号	成分参数			后验概率	判别相	原相号	判别相	后验概率	粒度参数			
		x_1	x_2	x_3						M_z	σ	S_{k1}	k_g
三角洲	1	1.50	0.01	0.33	0.351 4	1	1	1	0.810 1	3.47	0.59	0.39	1.40
	2	2.67	0.07	0.11	0.452 5	1	1	2	0.641 0	2.85	1.70	0.45	2.04
	3	1.33	0.30	0.01	0.951 5	2	1	3	0.436 7	4.70	1.56	0.43	1.46
	4	4.00	0.07	0.08	0.657 2	1	1	2	0.760 3	3.10	2.00	0.45	1.63
	5	1.50	0.05	0.33	0.547 7	1	1	1	0.846 0	3.55	0.53	0.47	1.55
	6	2.12	0.06	0.29	0.609 1	1	1	1	0.740 3	3.20	0.99	0.10	1.29
	7	1.60	0.10	0.19	0.468 9	3	1	1	0.8404	3.67	1.22	0.41	2.34
	8	2.60	0.08	0.19	0.548 1	1	1	1	0.974 9	4.13	0.67	0.26	1.91
	9	2.14	0.06	0.26	0.575 6	1	1	1	0.428 6	4.57	0.37	0.47	1.56
	10	1.25	0.05	0.35	0.528 5	1	1	1	0.919 6	3.48	1.20	0.30	2.46
	11	1.22	0.06	0.35	0.527 1	1	1	1	0.838 3	4.54	1.19	0.28	1.73
	12	6.14	0.10	0.40	0.978 0	1	1	1	0.879 5	3.69	0.89	0.67	2.61
	13	1.50	0.10	0.22	0.467 0	3	1	1	0.715 3	3.69	0.90	0.51	1.75
	14	3.00	0.35	0.01	0.901 7	2	1	1	0.964 6	3.34	0.81	0.43	2.61
	15	4.90	0.30	0.05	0.579 7	1	1	1	0.942 7	3.46	0.87	0.42	2.41
浊流	16	0.50	0.28	0.01	0.963 9	2	2	1	0.504 8	4.57	1.40	0.35	1.45
	17	2.33	0.26	0.09	0.688 5	2	2	2	0.661 7	3.21	1.51	0.54	1.73
	18	1.50	0.25	0.20	0.496 9	2	2	2	0.861 1	2.80	1.78	0.46	1.45
	19	0.25	0.16	0.06	0.814 3	2	2	2	0.550 9	3.95	1.60	0.35	1.31
	20	0.82	0.20	0.06	0.831 0	2	2	2	0.513 0	4.00	1.58	0.43	1.42
	21	1.97	0.20	0.02	0.781 0	2	2	2	0.574 9	3.88	2.21	0.64	1.38
	22	2.33	0.16	0.00	0.689 1	2	2	1	0.422 5	3.13	1.49	0.60	2.46
	23	1.97	0.20	0.01	0.800 7	2	2	2	0.912 2	1.57	1.00	0.35	1.20
	24	1.04	0.10	0.05	0.601 0	2	2	2	0.988 1	1.66	1.67	0.59	0.85
	25	2.33	0.22	0.07	0.647 0	2	2	3	0.449 2	4.39	1.62	0.56	1.37
	26	2.41	0.25	0.10	0.609 8	2	2	2	0.892 0	2.62	1.76	0.35	1.34
	27	1.50	0.18	0.11	0.577 8	2	2	3	0.653 2	4.73	2.02	0.54	1.39
	28	2.59	0.25	0.06	0.699 5	2	2	3	0.608 8	4.55	1.86	0.61	1.51
	29	1.00	0.45	0.01	0.992 7	2	2	2	0.818 0	2.55	1.85	0.41	1.82
	30	1.22	0.14	0.01	0.770 5	2	2	2	0.488 2	3.17	1.24	0.13	1.21
	31	2.41	0.25	0.04	0.776 6	2	2	2	0.567 0	3.57	1.46	0.62	1.73
	32	1.04	0.20	0.02	0.870 6	2	2	2	0.490 0	4.08	1.53	0.54	1.47
	33	1.67	0.19	0.06	0.710 8	2	2	2	0.796 0	3.23	2.07	0.48	1.36
风暴流	34	1.00	0.05	0.20	0.573 7	3	3	1	0.809 9	3.72	0.94	0.44	1.87
	35	2.03	0.05	0.40	0.708 7	1	3	3	0.854 0	4.60	2.24	0.78	2.24
	36	1.20	0.06	0.10	0.487 1	3	3	1	0.482 6	4.11	1.05	0.44	1.23
	37	2.00	0.14	0.14	0.366 7	3	3	3	0.880 7	5.54	2.31	0.64	1.34
	38	1.58	0.13	0.17	0.414 0	3	3	3	0.889 2	5.03	2.76	0.73	1.69
	39	1.06	0.08	0.27	0.518 6	3	3	2	0.602 3	4.34	1.96	0.57	0.91
	40	2.33	0.07	0.08	0.732 2	3	3	3	0.632 9	4.57	2.22	0.60	1.32
	41	1.30	0.15	0.20	0.408 8	3	3	3	0.689 5	4.25	2.21	0.63	1.91
	42	3.55	0.11	0.11	0.615 9	1	3	1	0.580 1	3.95	1.62	0.61	2.89
	43	1.50	0.15	0.30	0.528 8	1	3	3	0.486 0	4.45	1.77	0.43	1.94
	44	1.50	0.19	0.16	0.455 2	2	3	2	0.453 8	3.68	1.61	0.58	1.93
	45	2.33	0.13	0.12	0.374 0	1	3	2	0.800 2	3.22	1.74	0.27	1.21

$$F_1(\boldsymbol{X}) = 6.244\,2M_z + 4.474\,6\sigma - 4.019\,2S_{k1} + 11.018\,0k_g - 24.847\,2 \quad (三角洲相)$$

$$F_2(\boldsymbol{X}) = 4.979\,2M_z + 7.583\,4\sigma + 0.922\,1S_{k1} + 8.310\,4k_g - 22.013\,0 \quad (浊流相)$$

$$F_3(\boldsymbol{X}) = 6.419\,9M_z + 8.155\,1\sigma + 1.433\,5S_{k1} + 9.612\,4k_g - 31.312\,9 \quad (风暴流相)$$

对于上述判别函数来说，无论是对表 5-3 中样品的判别检验效果，还是在研究东濮凹陷西部沙三段沉积相时的使用情况，均反映出成分成熟度参数判别效果优于粒度参数判别效果。其原因在于粒度参数比成分成熟度参数更具多解性，即沉积物的粒度受沉积环境水动

力条件的控制，但不同的沉积环境可具有相似的水动力条件。由此表明，定量研究地质问题时，地质参数的选择是非常重要的，应该尽可能地选择那些地质含义明确、代表作用强的参数，否则，无论多么完善的数学方法也不能使地质问题得到满意的解释。

【例 4】预报油气勘探成功率。

四川盆地侏罗系自流井群大安寨组评价区划分为 675 个单元（参见回归分析应用实例中的【例 1】）。在 675 个单元中，有钻探资料的单元有 139 个，其中 57 个单元获得了工业油气井，把这些单元记为 A 组，其勘探成功率为 1；未获得工业油气井，经过研究认为也不可能获得工业油气井的单元有 38 个，把这些单元记为 B 组，其勘探成功率为 0。以 14 个地质变量对两组单元作逐步判别分析，取 $f_1=f_2=2$，得到勘探成功率为 1 的 A 组的判别函数 $F_A(\boldsymbol{X})$ 和勘探成功率为 0 的 B 组的判别函数 $F_B(\boldsymbol{X})$：

$$F_A(\boldsymbol{X}) = 1.3911\times10^{-2}x_2 + 3.8300\times10^{-3}x_3 + 1.6138\times10^{-2}x_4 + 5.2713\times10^{-1}x_5 + 8.9093\times10^{-2}x_7 - 34.9155$$

$$F_B(\boldsymbol{X}) = 1.7956\times10^{-2}x_2 + 7.6630\times10^{-3}x_3 + 2.1202\times10^{-2}x_4 + 2.9107\times10^{-1}x_5 + 3.2441\times10^{-2}x_7 - 37.3755$$

选入的变量 x_2, x_3, x_4 属于构造因素，变量 x_5 代表生储盖条件的搭配，而 x_7 代表生储条件。上述五个变量的组合反映了受岩性岩相控制的裂缝性油气藏的特征。

利用 $F_A(\boldsymbol{X})$，$F_B(\boldsymbol{X})$ 对 A，B 两组单元的回判结果为：A 组的 57 个单元判对了 54 个，B 组的 38 个单元判对了 36 个，对判率大于 94%，判别函数是高度显著的。

将评价单元对应的五个变量值代入 $F_A(\boldsymbol{X})$，计算出属于 A 组的后验概率，也就是单元的勘探成功率，在此基础上绘出勘探成功率预测等值线图（图 5-5）。结合其他资料，应在此图上选择成功率相对高的区块勘探。

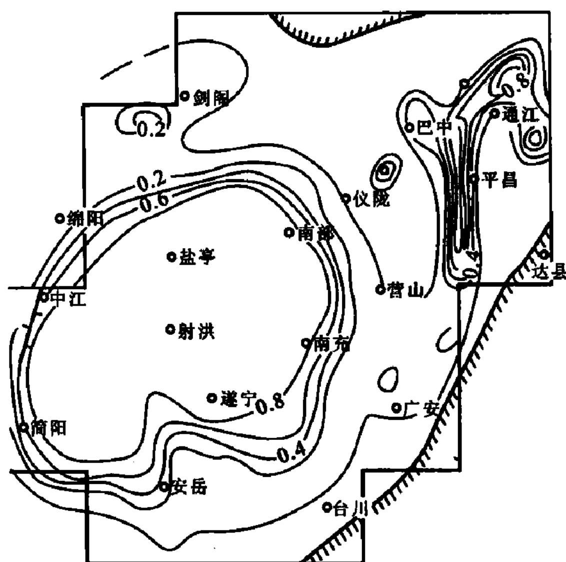


图 5-5　大安寨组勘探成功概率预报图

（据陈立平，陈子恩）

思考与练习

1. 什么是判别分析？
2. 试述建立线性判别函数的费歇尔准则。
3. 如何利用线性判别函数对样品所属的总体做出判别？
4. 试述 Bayes 准则下建立多总体判别一般判别函数的基本原理。
5. 逐步判别分析为何提出？
6. 试述逐步判别分析的基本思想。
7. 已知总体数 $G=4$，样品总数 $n=42$，逐步判别分析共引入变量数 $p=3$，计算出的统

计量 $F=2.5$，给定检验水平 $\alpha=0.05$，试检验判别函数的显著性。

8. 对于【例 3】中的成分成熟度和粒度数据，试用逐步判别分析程序求三角洲、浊流和风暴流沉积相的判别函数，给定检验水平 $\alpha=0.1$，检验所求判别函数的显著性，并确定取自上述沉积相的样品 $\boldsymbol{X}=(3.67,0.07,0.01)$，$\boldsymbol{Y}=(1.97,0.20,0.02)$，$\boldsymbol{Z}=(1.58,0.13,0.11)$ 的沉积相。

第六章 趋势面分析

任何一个地质变量 z，如地层面的埋藏深度、地层的厚度、储层中油气的粘度和比重、地层水的矿化度、油气地表地球化学勘探指标等，其观测值 z_i 与观测点的地理坐标 (x_i, y_i) 一起构成三维空间中的点，将其记为 $M_i(x_i, y_i, z_i)(i=1,2,\cdots,n)$，趋势面分析就是在这个点的控制下，拟合一个连续的数学曲面，以此研究地质变量在区域上和局部范围内变化规律的一种统计分析方法。拟合出的数学曲面叫做趋势面，表示地质变量的变化趋势。地质变量的观测值分布在趋势面上或者它的上下附近(图 6-1)。在趋势面分析中，多项式和傅里叶级数是常用的数学模型。

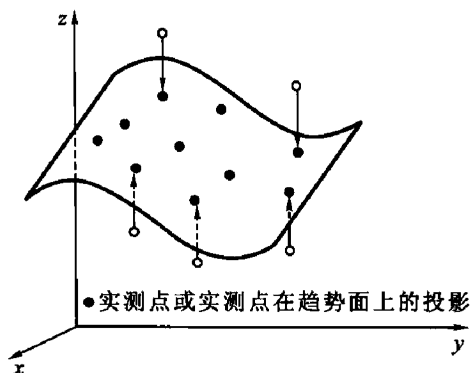


图 6-1 趋势面与观测值的关系示意图

§1 多项式趋势面分析

一、多项式曲面的一般形式

多项式曲面的一般形式为：

$$z = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2 + \cdots \tag{6-1}$$

式中 z——地质变量；

x, y——观测点的地理坐标；

$a_1, a_2, a_3, \cdots$——待定常数与系数。

如果式(6-1)中自变量的最高次数为 k，则称式(6-1)为 k 次多项式曲面。曲面的形态将随着 k 的增大而变得更加复杂(图 6-2)。

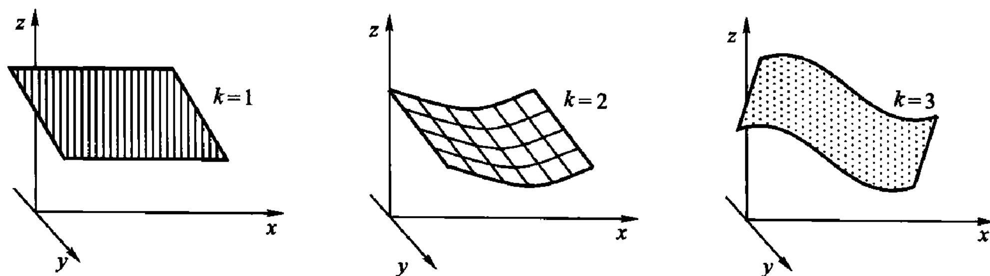


图 6-2 多项式趋势面示意图

二、多项式系数的个数及系数的确定

1. 多项式系数的个数

对于 k 次多项式，其系数与常数项的个数 $p=(k+1)(k+2)/2$。

2. 多项式系数的确定

具体地说，拟合多项式曲面就是根据点 $M_i(x_i, y_i, z_i)(i=1,2,\cdots,n)$，确定式(6-1)中的

系数和常数项。

设 $a_1,a_2,a_3\cdots$的估计值为 $b_1,b_2,b_3\cdots$，那么趋势多项式方程为：

$$\hat{z} = b_1 + b_2x + b_3y + b_4x^2 + b_5xy + b_6y^2 + \cdots \tag{6-2}$$

把观测点的地理坐标$(x_i,y_i)(i=1,2,\cdots,n)$代入式(6-2)，得观测点上地质变量的趋势值：

$$\hat{z}_i = b_1 + b_2x_i + b_3y_i + b_4x_i^2 + b_5x_iy_i + b_6y_i^2 + \cdots \tag{6-3}$$

如果统计量

$$Q_1 = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2$$

达到最小，表明观测值与趋势值最接近。由 Q_1 可知，它是关于 $b_1,b_2,b_3\cdots$的一个二次函数，并且 $Q_1>0$，根据极值原理有：

$$\frac{\partial Q_1}{\partial b_k} = 0 \qquad (k = 1,2,\cdots,p) \tag{6-4}$$

式中　p——多项式系数与常数项的个数。

从式(6-4)可以解出 $b_1,b_2,b_3\cdots$，得到趋势多项式方程式(6-2)。

对式(6-4)化简整理，写成矩阵形式有

$$\boldsymbol{AB} = C \tag{6-5}$$

并称式(6-5)为正规方程组，从中可以解出 $b_1,b_2,b_3\cdots$，得到趋势多项式方程式(6-2)。其中 $\boldsymbol{A}=\boldsymbol{X'X}$，$\boldsymbol{C}=\boldsymbol{X'Z}$，$\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)'$，$\boldsymbol{B}=(b_1,b_2,\cdots,b_p)'$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & y_1 & x_1^2 & x_1y_1 & y_1^2 & x_1^3 & x_1^2y_1 & x_1y_1^2 & y_1^3 & \cdots \\ 1 & x_2 & y_2 & x_2^2 & x_2y_2 & y_2^2 & x_2^3 & x_2^2y_2 & x_2y_2^2 & y_2^3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 1 & x_n & y_n & x_n^2 & x_ny_n & y_n^2 & x_n^3 & x_n^2y_n & x_ny_n^2 & y_n^3 & \cdots \end{pmatrix}$$

三、趋势面的拟合度

趋势面的拟合度是指观测点上的观测值与趋势值在总体上的逼近程度。如果记

$$Q = \sum_{i=1}^{n}(z_i - \bar{z})^2, \qquad Q_2 = \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2$$

定义趋势面的拟合度为：

$$C = (Q_2/Q) \times 100\%$$

四、趋势面偏差图

偏差是地质变量的观测值 z_i 与趋势值 $\hat{z}_i$ 之差，即 $\Delta z_i = z_i - \hat{z}_i$。趋势面偏差图是以偏差为绘图数据绘制的等值线图(图 6-3)。在偏差图上，偏差大于 0 的等值线圈出的区域称为正偏差区(正剩余区、正残差区)，偏差小于 0 的等值线圈出的区域称为负偏差区(负剩余区、负残差区)。关于正、负偏差的地质解释，要依据偏差的地质内涵进行具体分析。例如，若地质变量是地层面埋藏深度的绝对值，那么偏差图上的正偏差区在某种意义上是放大了的局部洼陷，而负

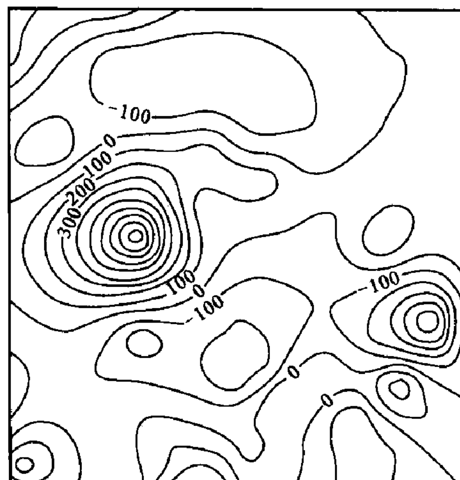


图 6-3　二次趋势面偏差等值线图

偏差区则意味着局部突起。分析这种偏差对研究盆地的油气分布和查找隐蔽圈闭是非常有益的。

五、趋势面异常图

通常来说，任何地质变量 z 的观测值 z_i 由区域性趋势（背景）分量 u_i、局部特征（局部异常）分量 v_i 和随机干扰分量 r_i 三部分组成，即：

$$z_i = u_i + v_i + r_i$$

式中 u_i 就是趋势值 $\hat{z}_i$，而 $v_i + r_i$ 则是偏差 Δz_i。由此可知，若直接由偏差来研究、分析变量 z 的局部特征，则会受到 r_i 的影响，降低研究结果的可靠性。所以，在进行偏差分析时，最好把 Δz_i 中包含的 r_i 消除或者对其进行抑制。在实际计算中，常取 m 个正偏差 Δz_i^+ 的平均值

$$e = \frac{1}{m}\sum_{i=1}^{m}\Delta z_i^+$$

或者标准差

$$\sigma = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\Delta z_i^+ - e)^2}$$

的 2 倍作为随机分量的估计值，并称其为异常限。

从正偏差 Δz_i^+ 中划分正异常时，异常限为正值，并称其为异常下限。异常下限确定后，称 $\Delta z_i^+ > e$（或 2σ）的点为正异常点，该点的异常值为：

$$v_i = \Delta z_i^- - e \text{ 或者 } v_i = \Delta z_i^+ - 2\sigma$$

从负偏差 Δz_i^- 中划分负异常时，异常限为负值，并称其为异常上限。异常上限确定后，称 $\Delta z_i^- < -e$（或 -2σ）的点为负异常点，该点的异常值为：

$$v_i = \Delta z_i^- + e \text{ 或者 } v_i = \Delta z_i^- + 2\sigma$$

异常点确定后，根据异常下、上限可以在偏差图上圈出正、负异常区，也可用异常点的异常值绘制等值线图，并把该等值线图称为趋势面异常图。

值得注意的是：这里的异常限是一个统计估计值，并非各点上的随机分量值。因此，它是一个可以改变的量。在实际工作中，可以根据资料的实际情况和要求，改变它的大小，使异常和随机分量得到最佳的分离。

六、关于趋势面的次数

利用趋势面分析的方法研究地质变量在区域上和局部范围内的变化规律时，究竟采用多少次的多项式合适？即多项式的次数 k 取多大为宜？

确定 k 的方法如下：

方法 1：对地质变量进行一次、两次、……趋势面分析，相应的拟合度为 $C_1, C_2, \cdots$，在 k-C 坐标系内作 k, C 的散点图（图 6-4）。连接各点形成一条折线，在折线上取斜率最大的线段对应的 k 作为趋势面的次数。

方法 2：对地质变量进行一次、两次、……趋势面分析，相应的拟合度为 $C_1, C_2, \cdots$。对预先给定的小正数 ε，当 $C_{i+1} - C_i < \varepsilon$ 时，取 C_i 对应的 k 作为趋势面的次数。

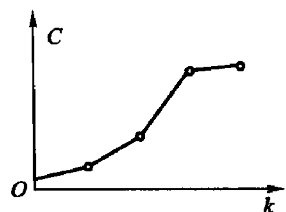


图 6-4　选择 k 示意图

§2 调和趋势面分析

调和趋势面分析的方法原理及研究过程与多项式趋势面相同，不同之处仅是它采用的数学模型为傅里叶级数。因此，本节在多项式趋势面分析的基础上，仅作如下说明。

一、正弦波的调和与叠加

1. 正弦波的调和

正弦波

$$z = a\sin(\omega x + \varphi) \tag{6-6}$$

是最简单的波(图 6-5)，其波长 $\lambda = 2\pi/\omega$。

式中 φ——初相位；

ω——角频率；

a——振幅；

x——自变量。

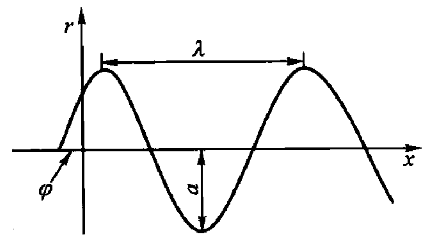


图 6-5 正弦波示意图

展开式(6-6)得：

$$\begin{aligned} z &= a\sin(\omega x + \varphi) = a\sin\varphi\cos\omega x + a\cos\varphi\sin\omega x \\ &= A\cos\omega x + B\sin\omega x \end{aligned} \tag{6-7}$$

由式(6-7)可知：单一的正弦波可以分解为正弦波与余弦波的和，反之可写为单一的正弦波。

一维调和是式(6-7)给出的正弦波，当 $\omega=2k\pi/\lambda$ 时，即

$$z_k = a_k\sin(2k\pi x/\lambda)$$

时称为一维 k 阶调和，它的波长等于 λ/k。

二维调和是以 x,y 为自变量的正弦(余弦)函数的积，即

$$a\cos\omega x\cos vy,\quad b\sin\omega x\cos vy,\quad c\cos\omega x\sin vy,\quad d\sin\omega x\sin vy \tag{6-8}$$

当式(6-8)中 $\omega=2m\pi/\lambda_1, v=2n\pi/\lambda_2$ 时，称式(6-8)为二维 m，n 阶调和。

2. 正弦波的叠加

把振幅相同或不同的一维调和叠加起来，则可形成各种复杂的曲线。同理，多个简单二维调和的叠加，便构造出形态复杂的曲面(图 6-6)。

二、调和趋势面

二元傅里叶级数是二维调和的线性组合，它的一般形式为：

$$\begin{aligned} z = F(x,y) = \sum_{t=0}^{r}\sum_{k=0}^{s}[&E_{tk}\cos(2t\pi x/L)\cos(2k\pi y/H) + C_{tk}\sin(2t\pi x/L)\cos(2k\pi y/H) \\ &+ P_{tk}\cos(2t\pi x/L)\sin(2k\pi y/H) + W_{tk}\sin(2t\pi x/L)\sin(2k\pi y/H)] \end{aligned} \tag{6-9}$$

拟合调和趋势面就是根据点 $M_i(x_i,y_i,z_i)(i=1,2,\cdots,n)$，按最小二乘法原理，确定式(6-9)中 $E_{tk},C_{tk},P_{tk},W_{tk}$ 的估计值 $a_{tk},b_{tk},c_{tk},d_{tk}$，得到方程：

$$\begin{aligned} \hat{z} = F(x,y) = \sum_{t=0}^{r}\sum_{k=0}^{s}[&a_{tk}\cos(2t\pi x/L)\cos(2k\pi y/H) + b_{tk}\sin(2t\pi x/L)\cos(2k\pi y/H) \\ &+ c_{tk}\cos(2t\pi x/L)\sin(2k\pi y/H) + d_{tk}\sin(2t\pi x/L)\sin(2k\pi y/H)] \end{aligned} \tag{6-10}$$

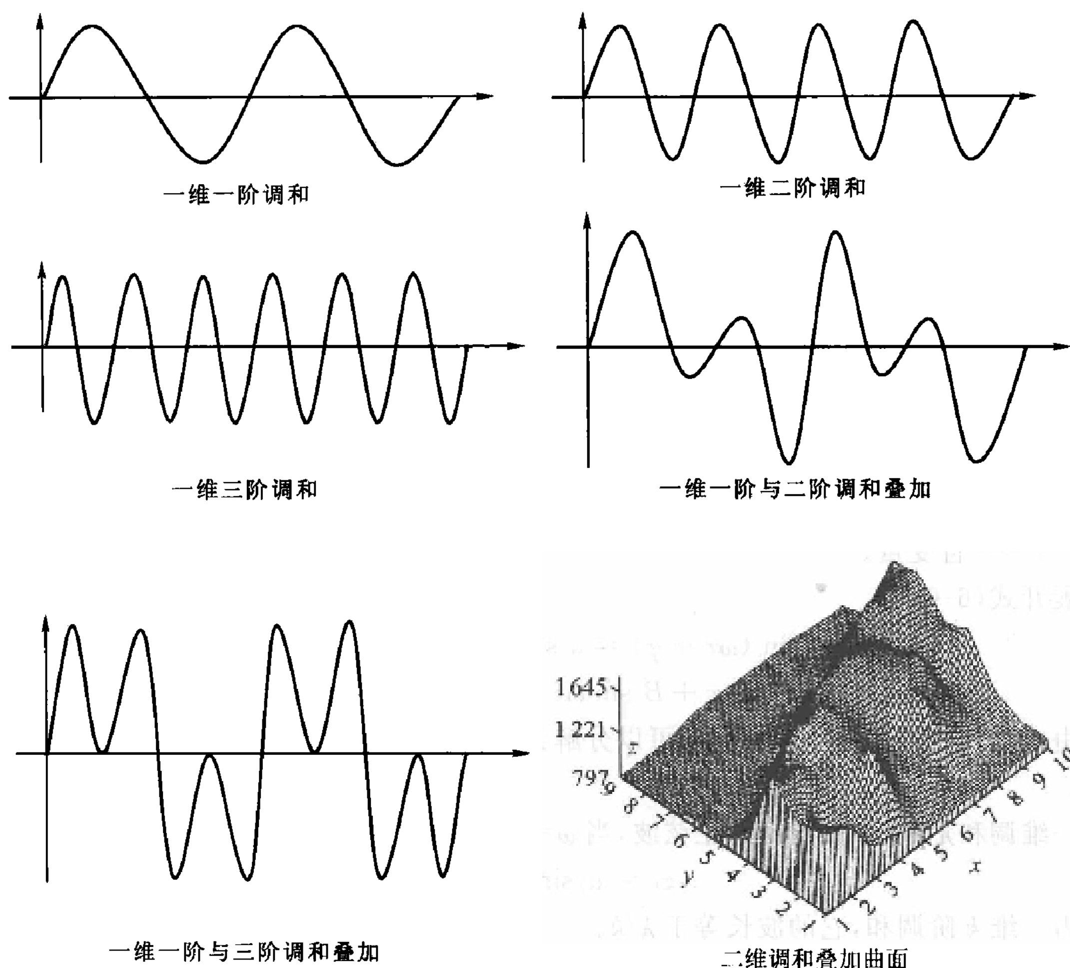


图 6-6 正弦波的叠加曲线与曲面

式中 $\hat{z}$——傅里叶级数趋势值；

r——x 方向上傅里叶级数的最大调和阶数；

s——y 方向上傅里叶级数的最大调和阶数；

L——x 方向的取样长度，即原图的横向长度；

H——y 方向的取样长度，即原图的纵向长度。

由式(6-10)可知，调和趋势面分析不仅具备多项式趋势面分析的功能，而且具有明显的波动特征。因此，它可以更有效地分离那些呈现出周期性不严格的地质变量(如地层面的波状起伏、沉积旋回、地球磁场的变化等)的趋势和异常，进而研究地质变量的波动特征。

§3 应用实例

【例 1】寻找有利油气储集构造。

油气田勘探实践表明，受构造因素控制的油气藏占有相当大的比重。但是，采用传统的地质方法研究构造与油气藏的关系时，有些局部构造，特别是低幅度的局部构造常常被区域性构造的展布特性所掩盖。趋势面分析方法能够突出局部异常，为寻找油气田提供新的参考依据。例如美国堪萨斯州东部的密西西比砾岩，其构造为一区域性的向西倾斜的单斜，其

上最大的局部圈闭高度较小，似乎不会形成大的油气藏。在地层面顶面构造二次趋势面异常图(图 6-7)上，存在大面积的正异常区，并且已探明的洛斯特普林油田基本上分布在正异常区内(图 6-8)。

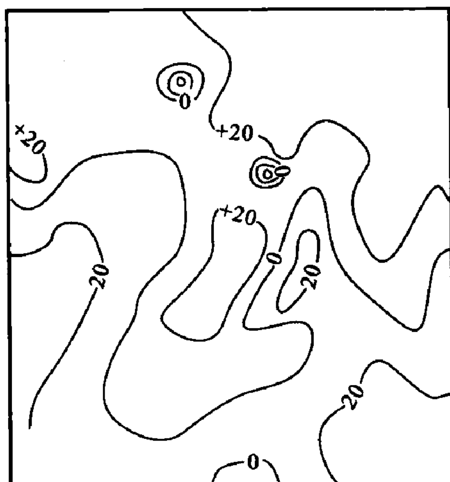


图 6-7 二次趋势面异常图
(据 Merriam 和 Harbaugh, 1964)

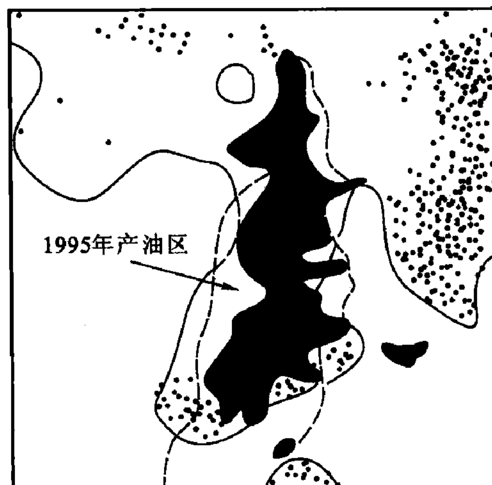
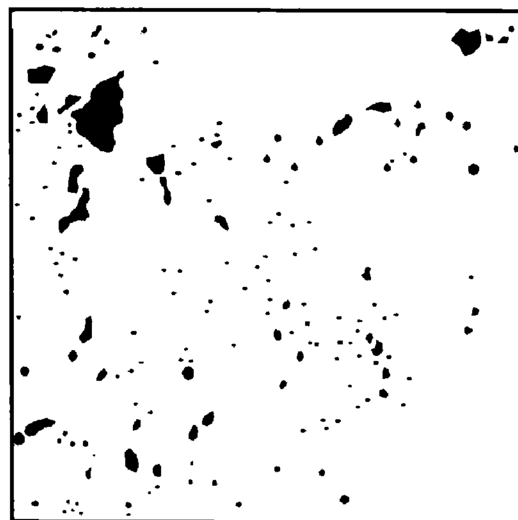


图 6-8 洛斯特普林油田(黑色)
(据 Merriam 和 Harbaugh, 1964)

图 6-9 是另一个用构造趋势面异常研究堪萨斯州东南地区(地区Ⅱ)正异常与油气田分布关系的实例，大多数油气田位于正异常区内。尽管油气的聚集与多种因素有关，但趋势面异常图仍可为油气勘探提供一些参考信息，如指出局部的特别是低幅度构造油气藏的潜在地区。值得注意的是：不能仅凭着趋势面分析结果而做出地质结论，而应该综合其他地质资料进行全面的地质解释，既要重视正异常区的有利部位，也不能忽视负异常区的有利地带，因为在负异常区可能找到岩性和地层油气藏。

(a) 异常等值线图



(b) 鞋带状砂岩油气田分布图

图 6-9 勘萨斯州东南地区(地区Ⅱ)构造二次趋势异常和油气田分布图
(据 Merriam 和 Harbaugh, 1964)

【例 2】寻找岩性-构造油气藏。

酒泉盆地北部单斜带经过多年勘探，在第三系火烧沟群先后发现了白杨河、单北和白东三个油田。在分析认识已有资料的基础上，试图利用趋势面分析方法在该单斜构造上寻找岩性-构造油气藏。

火烧沟群构造为一由北向南西倾斜的平缓单斜，其上有少数几个鼻状构造和膝状挠曲。在一次趋势面图上，该群顶部构造背景为一南西 12°倾斜的平面，倾角为 11°6′。在构造异常图（图 6-10）上明显看出，已探明的油田全部位于正异常区，说明构造因素在控制油气上占有重要地位。但是白杨河与白东油田并不在正异常区的最高部位，而是在斜坡上或靠近正异常的边界处，这说明它们除受构造因素控制外，还受到岩性变化的制约（油层上倾方向物性变差），是典型的岩性-构造油气藏。由此得出：正异常区地层上倾方向的低渗透带是有利的构造部位。基于上述认识，提出几块有利面积，经初步勘探，在 A，B，E 三块面积内发现了好的油砂或工业性油流。

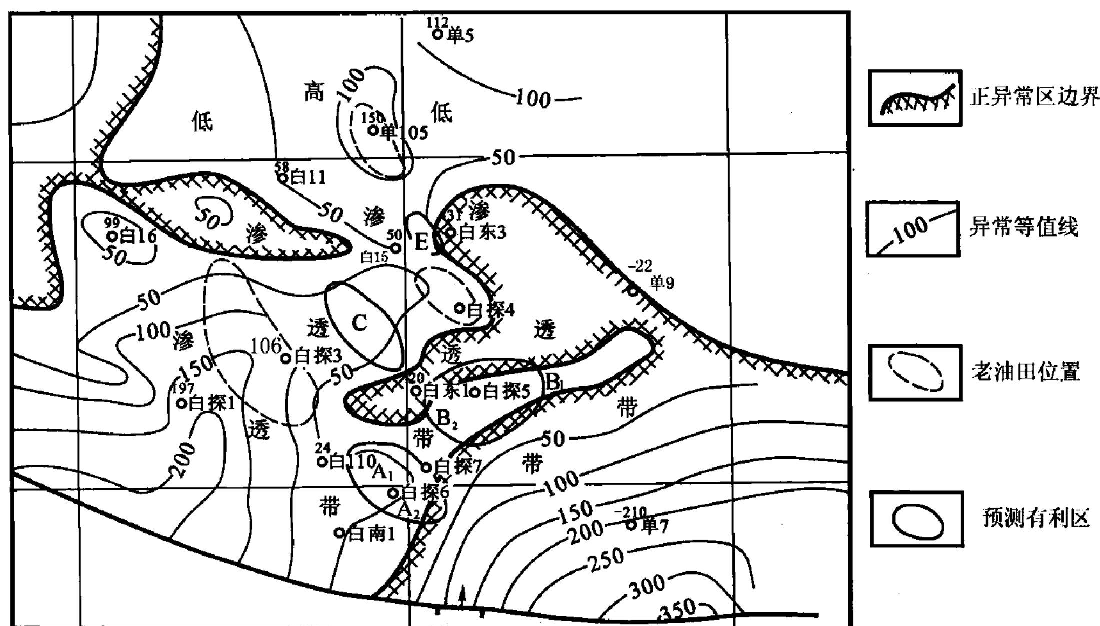


图 6-10　酒泉盆地北部单斜带第三系火烧沟群顶
一次趋势面异常等值图和有利区预测图
（据陈立官主编《油气田地下地质学》，1983）

【例 3】研究地下断裂分布。

把局部异常和随机干扰从数据中分离出来，以期从中找出隐蔽的或者被掩盖而又有意义的地质信息。岩体顶面的起伏虽有较大的随机性，但往往是连续变化的，然而构造断裂所造成的起伏则有线状分布、方向性明显和非连续的特点。这些特点是地下断裂能够通过趋势面分析加以显示的前提条件。

由于趋势图描绘了大范围的总体变化，而偏差图则包含了小范围的局部特征和无规律的随机成分。因此，对偏差值再进行分解，可得到次一级的趋势值和偏差值。因为第一次趋势图已将背景分离出去，于是第二次处理的数据基本上不包含主要的区域性分量了，尤其是在拟合度偏高的情况下更是如此。因此，第二次的趋势图表现了小范围的变化特征，而偏差

图则更集中地反映了更小的局部特征和随机成分。呈线状分布，且有一定方向性的断裂必然要在第二次趋势面偏差图中反映出来。

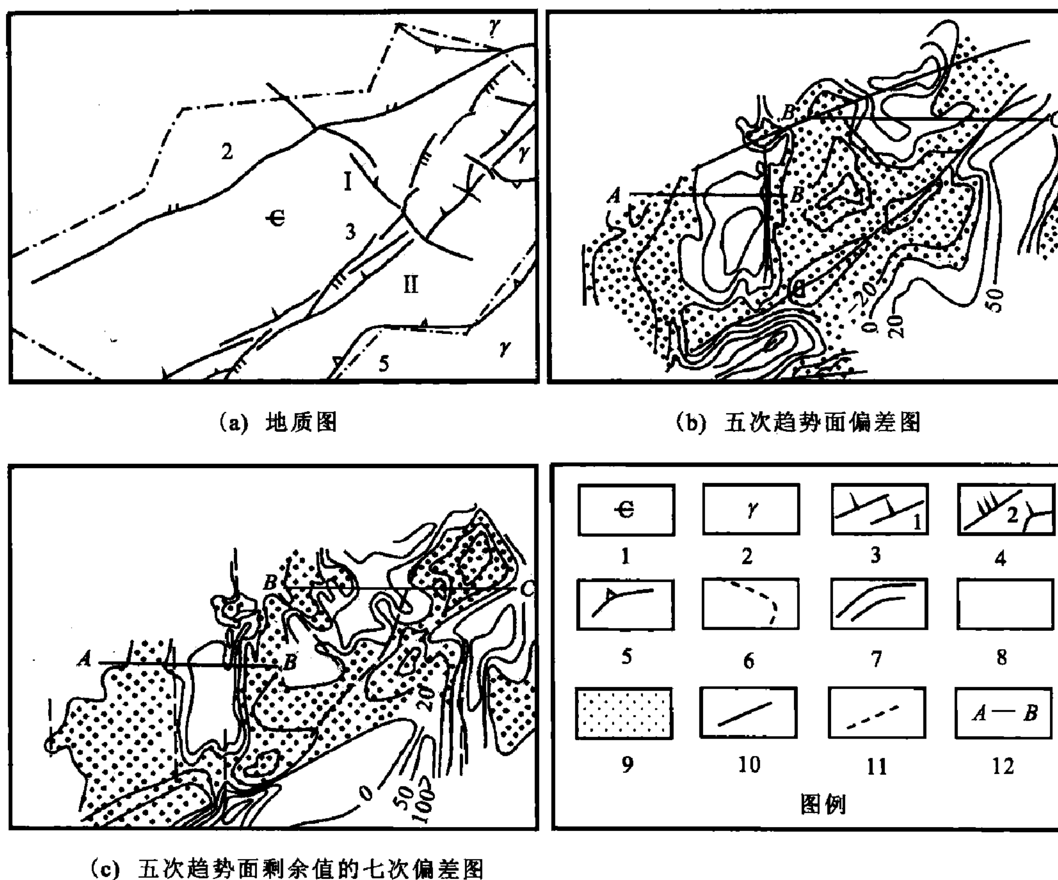


图 6-11 某研究区地质及趋势面偏差图

1—寒武纪地层；2—花岗岩；3—张性、张扭性断裂及编号；4—压性、压扭性断裂及编号；
5—岩体与围岩接触界线；6—数据点分布范围；7—偏差等值线；8—正偏差区；9—负偏差区；
10—已知断裂；11—趋势面分析预测断裂；12—剖面线位置及编号
（据陈立官主编《油气田地下地质学》，1983）

在据某地区寒武系地面资料绘制的断裂分布图（图 6-11a）和该区五次趋势面偏差图（图 6-11b）上，可以明显看出区内两条 NEE 走向的大断裂非常吻合。但图 6-11b 上最醒目之处是中央的 SN 向正偏差区等值线梯度变化特别大，使人们毫不怀疑这里潜藏着一条向东倾斜的断裂，但它在地面地质图上却没有反映。后经钻探证实，该区深部确实存在这条较大的断裂。最富有说服力和令人感兴趣的是五次趋势面偏差值的七次偏差图（图 6-11c，它是对五次趋势面偏差值又进行七次趋势面分析后作的七次偏差图），从已知断裂部位等值线的方向性，梯度变化，正、负偏差值的界线看，都非常吻合，甚至有些部位把呈平行带状出现的断裂更准确、更细致地反映出来了。从此例不难看出，趋势面分析对断裂构造的研究是一种很有效的方法。一般情况下，在趋势面偏差图上，等值线排列的方向性、规模大小和梯度的变化反映地下断裂及其产状，尤其对具有一定规模和垂向位移较大的断裂，效果会更好。

胜利油田地质科学研究院对标准层标高进行趋势面分析，在趋势面正偏差区查找低幅度构造圈闭，增加了石油地质储量，同时指出：偏差图上的等值线密集带，往往是断裂分布的

位置,密集等值线的方向指示了断层的延伸方向。

【例 4】预测有利勘探区。

利用地震波通过含油气层时高频成分被强烈吸收,而低频能量相应增强的特征反演预测有利含油区的方法称为 HCI(碳氢检测)技术。这种技术可提供多种资料,碳氢检测地震特征图(图 6-12)和碳氢检测地震综合图(图 6-13)是其中的两种。

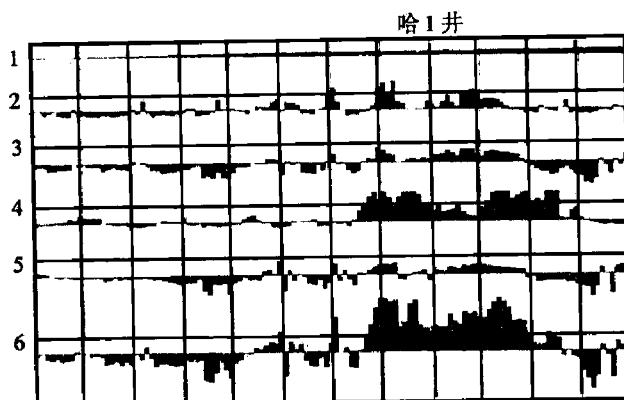


图 6-12　碳氢检测地震特征图

地震特征图由速度变化特征、10 Hz 宽带能量百分比、10 Hz 能量、平均频率特征、宽带能量特征(能谱)及峰值频率特征等六种地震信息组成。大量资料表明,在地震特征图上,有利含油气层段明显地反映出 10 Hz 宽带能量百分比、10 Hz 能量、宽带能量特征等指标(图 6-12 中的 2,3,5)增强,速度变化、平均频率及峰值频率特征(图 6-12 中的 1,4,6)降低,而总值升高。

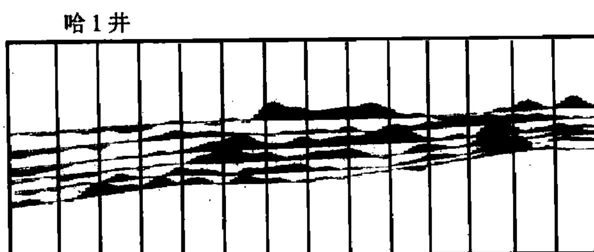


图 6-13　碳氢检测地震综合图

对勘探目的层及其顶和底进行 HCI 处理,获得储层各自的地震特征图后,再将各层的总值绘制在一张图上,即形成碳氢检测地震综合图。

1982 年,石油地球物理勘探局第四勘探公司在哈南地区选用了四条 HCI 剖面,对 400 个观测点的地震综合信息进行计算,反映上下中三层目的层检测效果,绘制了哈南地区 HCI 三次趋势面偏差值的四次趋势面图(图 6-14),图上显示出两个正偏差区。东区呈 NNE 走向,是一个以正值>4 的南北两高点的偏差区,自高点向翼部(短轴方向)呈每千米递减 4 的比率。西区整体亦呈 *NE* 向,面积约为东区的两倍,主体在北部,是一个高点正值>4、圈闭主体向北开口的正偏差区,短轴方向的递减比率与东区相同。它的特点是向西南及东部均有平缓延伸的正偏差区。结合其他资料,拟定东西两个正偏差区为有利勘探区。

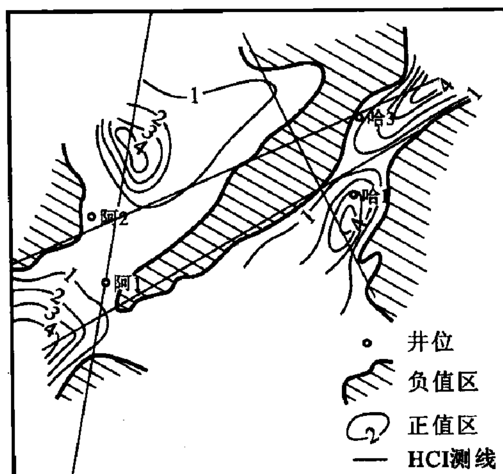


图 6-14　哈南地区 HCI 三次趋势面偏差值的四次趋势面图

1984 年以来在相应地区经过地震详查及钻探,发现了哈南油田和蒙古林油田。在东区发现了一个 NE 走向,由夫特西、哈南、布敦北、额汉南四个构造组成、面积为 70 km^2 的潜山背斜构造带。在哈南构造上的哈 1 井、哈 8 井均已见到自喷油流。在夫特西构造上钻探见到轻质含油砂

岩，表明是一个复式油藏。在西区主体部位发现了一个多高点背斜构造，经地震和钻探证实是一个具有较大面积、多套储层、较多地质储量的复式油藏。

思考与练习

1. 简述趋势面分析的概念及研究对象。
2. 简述多项式趋势面与调和趋势面在应用上的异同。
3. 试述求趋势面方程的方法原理。
4. 如何选择趋势面的拟合度？
5. 如何绘制趋势面偏差图和异常图？
6. 如何解释趋势面偏差图和异常图的地质含义？
7. 熟悉二维一次、二次、三次多项式趋势面模型。
8. 趋势面分析与回归分析在方法原理及应用方面有何异同？
9. 当趋势面分析结果出现地质变量不可能出现的结果时，称此时的趋势面发生畸变。试分析导致趋势面发生畸变的原因。

10. 假设某铜矿床的矿源层是下寒武系底部的黑色页岩层，化探找矿工作中已对黑色页岩层等间距网格采样，并分析了样品的铜元素含量(表 6-1)。试对铜元素含量作一次趋势面分析，并据分析结果预测铜矿的有利成矿区。

要求如下：

① 绘出一次趋势面图(绘 5 条等值线)；

② 绘出一次趋势面偏差图；

③ 绘出异常图，标示成矿有利区；

④ 如果更换采样点坐标(表 6-2)，对趋势面分析结果是否有影响？为什么？

表 6-1　采样点坐标及样品铜元素含量 单位：%

y \ x	−2	−1	0	1	2
2	6	6.5	5	2	3
1	6.5	8	6	3	2
0	5	6	6.5	4	3
−1	2	3	4	4	3
−2	3	2	3	3	4

表 6-2　采样点坐标及样品铜元素含量 单位：%

y \ x	10	11	12	13	14
14	6	6.5	5	2	3
13	6.5	8	6	3	2
12	5	6	6.5	4	3
11	2	3	4	4	3
10	3	2	3	3	4

第七章 因子分析

§1 因子分析概述

一、主因子

假设有 n 个样品，每个样品有 m 个变量，它们的原始观测值记为数据矩阵：

$$\boldsymbol{X}_{m\times n}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

当 $m=2$ 时，在 x_1,x_2 坐标系内做样品的散点图(图 7-1)。由此图看出：在 x_1,x_2 坐标系内，两个变量的方差差别不大，也就是说，每个变量所提供的信息都不可忽视；将 x_1,x_2 坐标系旋转为 f_1,f_2 新坐标系，在新坐标系内，两个变量的方差却有明显的差别，f_1 的方差占了总方差的绝大部分。由此表明，变量 f_1 能够反映原始数据所包含的绝大部分信息。根据坐标旋转变换，上述变量间的关系为：

$$\begin{cases} f_1=a_{11}x_1+a_{21}x_2 \\ f_2=a_{12}x_1+a_{22}x_2 \end{cases} \quad 或 \quad \begin{cases} x_1=a_{11}f_1+a_{12}f_2 \\ x_2=a_{21}f_1+a_{22}f_2 \end{cases} \tag{7-1}$$

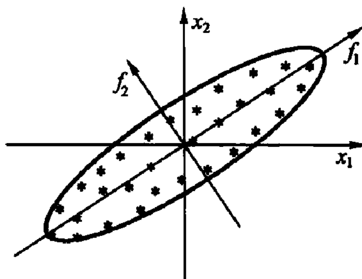


图 7-1 主因子示意图

当 f_2 的方差小于允许误差时，它就失去了存在的必要，此时就可用 f_1 代替两个原始变量，而式(7-1)改为：

$$f_i=a_{1i}x_1+a_{2i}x_2 \qquad 或 \qquad x_i=a_{i1}f_1+\alpha_i e_i \qquad (i=1,2) \tag{7-2}$$

其中 e_i 服从均值为 0，方差为 σ_i^2 的正态分布。

如果每个样品有 m 个变量，仿照式(7-1)可得：

$$f_i=a_{1i}x_1+a_{2i}x_2+\cdots+a_{mi}x_m \text{ 或 } x_i=a_{i1}f_1+a_{i2}f_2+\cdots+a_{im}f_m (i=1,2,\cdots,m) \tag{7-3}$$

由坐标旋转可知，式(7-3)中系数的平方和满足：

$$\sum_{j=1}^{m}a_{ji}^2=1 \qquad (i=1,2,\cdots,m)$$

在上述条件下，由原始变量经线性组合而得到的新变量 f_i 叫做综合变量或主因子。

把 m 个原始变量表示为 p 个主因子的线性组合，当 $p<m$，特别是 $p=2$ 时，就可以把高维空间中研究的问题降到二维空间中分析，从而化简研究系统。

二、因子分析

因子分析是研究变量间相关性、样品间相似性、两者成因联系以及探索它们之间产生上述关系的内因的一些多元统计分析方法的总称。因子分析的任务之一就是找出 $p(p<m)$ 个主因子，以此化简研究系统。根据研究对象，因子分析大致可分 R 型、Q 型因子分析和对应分析。

1. R 型因子分析

R 型因子分析是从研究 m 个变量的相关（相似）系数矩阵

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}$$

的内部结构出发，找出控制变量相关性的主因子 $f_i(i=1,2,\cdots,m)$，把变量 x_i 表示为 f_i 的线性组合，即：

$$x_i=a_{i1}f_1+a_{i2}f_2+\cdots+a_{im}f_m \qquad (i=1,2,\cdots,m) \qquad (7\text{-}4)$$

其中

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

叫做 R 型因子载荷矩阵。

在进行综合地质研究时，如果用前 $p(p\ll m)$ 个主因子就能解释原始数据 80%～90% 以上的信息，那么式(7-4)可改写为：

$$x_i=a_{i1}f_1+a_{i2}f_2+\cdots+a_{ip}f_p+\alpha_i e_i \quad (i=1,2,\cdots,m) \qquad (7\text{-}5)$$

上式表明，当 $p\ll m$，特别是当 $p=2$ 时，将极大地化简研究系统（在二维空间中分析研究高维空间中的问题），可更进一步探索变量的成因联系及其空间变化规律的控制因素。通常称式(7-5)为 R 型因子分析模型。

2. Q 型因子分析

Q 型因子分析是从研究 n 个样品的相似（相关）系数矩阵

$$\boldsymbol{Q}=\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix}$$

的内部结构出发，找出控制样品相似性的主因子 $f_j(j=1,2,\cdots,n)$，把样品 x_j 表示为 f_j 的线性组合，即：

$$x_j=a_{j1}f_1+a_{j2}f_2+\cdots+a_{jn}f_n \qquad (j=1,2,\cdots,n) \qquad (7\text{-}6)$$

其中

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

叫做Q型因子载荷矩阵。

若用前 $p(p \ll n)$ 个主因子化简样品研究系统，则有：

$$x_j = a_{j1}f_1 + a_{j2}f_2 \cdots + a_{jp}f_p + a_j e_j \qquad (j = 1,2,\cdots,n) \tag{7-7}$$

通常称式(7-7)为Q型因子分析模型。

3. 公因子

在因子分析模型中，各变量(或样品)中公有的因子 $f_1, f_2, \cdots, f_p$ 叫做公因子，它们是相互独立的理论变量，可将其理解为 p 维空间中相互垂直的 p 个坐标轴。单一变量(或样品)中特有的 $e_i(e_j)$ 叫做特殊因子，它们之间以及它们与所有公因子之间都是相互独立的。

因子载荷矩阵中的 a_{ij} 是第 i 个变量(或样品)在 f_j 轴上的负荷。

4. 对应分析

对应分析是把上述两种因子分析方法结合起来，在同一个空间里研究样品的相似性、变量的成因以及它们的分布规律，便于进行地质解释的一种多元统计分析方法。

由上述可知，进行因子分析的关键是确定因子分析模型，即确定主因子载荷矩阵。

§2 主因子载荷矩阵

主因子载荷矩阵又叫主因子的解。在此不进行理论上的阐述，仅从应用的角度考虑，给出因子模型中的主因子载荷矩阵。

一、R型主因子载荷矩阵及主因子个数

1. 主因子载荷矩阵

在R型因子分析中，设变量的相关系数矩阵 $\boldsymbol{R}$ 的特征值为 $\gamma_1, \gamma_2, \cdots, \gamma_m$，并且满足 $\gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_{p+1} \geqslant \cdots \geqslant \gamma_m$，这些特征值对应的特征向量为 $V_1, V_2, \cdots, V_{p+1}, \cdots, V_m$。若选定前 p 个主因子，令

$$\lambda_i = \gamma_{i+1}(i = 1,2,\cdots,p)$$
$$U_i = V_{i+1}(i = 1,2,\cdots,p)$$

那么可以证明前 p 个主因子载荷矩阵为：

$$\boldsymbol{A}_1 = (a_{ij})_{m\times p} = (u_{ij}\sqrt{\lambda_j})_{m\times p}$$

相应的R型因子分析模型为式(7-5)。

由上述可知，如果R型因子分析选 p 个主因子，那么 A_1 是相关系数矩阵 $\boldsymbol{R}$ 的特征值 $\gamma_{j+1}(j=1,2,\cdots,p)$ 的平方根与对应特征向量 $V_{j+1}(j=1,2,\cdots,p)$ 的积。

2. 主因子个数

对于主因子来说，各主因子的方差贡献 S_j^2 等于对应的特征值，即：

$$S_j^2 = \sum_{i=1}^{m} a_{ij}^2 = \lambda_j$$

用JACOBI法可求出相关(相似)系数矩阵的全部特征值和单位特征向量。若 m 个特征值从大到小排列为 $\lambda_1, \lambda_2, \cdots, \lambda_m$，计算特征值累计百分比

$$C=\sum_{j=1}^{p}\lambda_j\Big/\sum_{i=1}^{m}\lambda_i\times 100\% \qquad (p\leqslant m)$$

然后根据 C 确定主因子数 p。

二、Q 型主因子载荷矩阵及主因子个数

1. 主因子载荷矩阵

在 Q 型因子分析中，设样品的相似系数矩阵 Q 的特征值为 $\gamma_1,\gamma_2,\cdots,\gamma_n$，并且满足 $\gamma_1\geqslant\gamma_2\geqslant\cdots\geqslant\gamma_{p+1}\geqslant\cdots\geqslant\gamma_n$，这些特征值对应的特征向量为 $V_1,V_2,\cdots,V_{p+1},\cdots,V_n$。若选定前 p 个主因子，令

$$\lambda_j=\gamma_{j+1} \qquad (j=1,2,\cdots,p)$$

$$U_i=V_{i+1} \qquad (i=1,2,\cdots,p)$$

那么可以证明前 p 个主因子的主因子载荷矩阵为：

$$\boldsymbol{A}_1=(a_{ij})_{n\times p}=(u_{ij}\sqrt{\lambda_j})_{n\times p}$$

相应的 Q 型因子分析模型为式(7-7)。

2. 主因子个数

主因子个数的确定方法参见 R 型因子分析，注意：

$$C=\sum_{j=1}^{p}\lambda_j\Big/\sum_{i=1}^{n}\lambda_i\times 100\% \quad (p\leqslant n)$$

§3 方差最大正交旋转

一、R 型因子分析方差最大正交旋转

确定了 p 个主因子，即确定了 p 维空间，此时还要对主因子轴进行旋转(图 7-2)。对于标准化变量来说，主因子轴旋转的目的是使第 j 个公因子的代表性变量 x_i 在 f_j 轴上的系数等于或者趋近于 1，而在其他公因子轴上的系数等于或者趋近于 0，以便于解释主因子的地质意义。

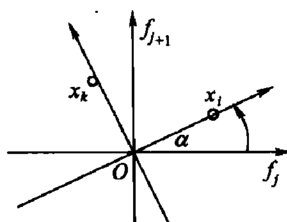


图 7-2 因子旋转示意图

1. 方差最大正交旋转

实现上述目的，从数学上讲，就是对主因子载荷矩阵 $\boldsymbol{A}_1$ 实施正交变换。常用的旋转方法是方差最大正交旋转。这种旋转方法使 p 个主因子保持彼此正交，并且使因子载荷矩阵中的各因子载荷平方后的方差达到最大。

对于 R 型因子载荷矩阵 $\boldsymbol{A}_1$ 来说，对因子 f_j 的简化效果，可用因子载荷平方的方差

$$V_j=\frac{1}{m}\sum_{i=1}^{m}(b_{ij}^2-\frac{1}{m}\sum_{i=1}^{m}b_{ij}^2)^2$$

来描述。其中 b_{ij} 是 $\boldsymbol{A}_1$ 经过正交旋转后所得到的因子载荷矩阵 $\boldsymbol{B}$ 中的元素。为避免出现负

值，故取 b_{ij}^2。如果 V_j 达到最大，则因子 f_j 得到了最优的简化。此时，公因子 f_j 的代表性变量在 f_j 轴上的系数等于或者趋近于 1，而在其他公因子轴上的系数等于或者趋近于 0。

对 p 个因子的简化效果可用因子载荷平方的方差之和

$$V = \sum_{j=1}^{p} V_j = \sum_{j=1}^{p} \frac{1}{m} \sum_{i=1}^{m} \left(b_{ij}^2 - \frac{1}{m} \sum_{i=1}^{m} b_{ij}^2 \right)^2 \tag{7-8}$$

来衡量。当 V 达到最大时，p 个因子都得到了最优简化。

考虑到诸公因子对变量总方差贡献的差异，将式(7-8)中的 b_{ij}^2 除以 h_i^2，并对该式两边乘以 m 化简，记为：

$$V^{\cdot} = m \sum_{j=1}^{p} \sum_{i=1}^{m} (b_{ij}^2 / h_i^2)^2 - \sum_{j=1}^{p} \left(\sum_{i=1}^{m} b_{ij}^2 / h_i^2 \right)^2 \tag{7-9}$$

式中　h_i^2——全部公因子对变量 x_i 的总方差所作的贡献，$h_i^2 = \sum_{j=1}^{p} a_{ij}^2$。

要求 $V^{\cdot}$ 达到最大，问题就归结为求一个 $p \times p$ 阶的正交矩阵 $\boldsymbol{T}$，使 $\boldsymbol{B} = \boldsymbol{A}_1 \boldsymbol{T}$ 能满足 $V^{\cdot}$ 为最大的条件。

对于 f_g, f_q 因子平面，可以对 $\boldsymbol{A}_1$ 作如下正交变换：

$$\boldsymbol{T}_{gq} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & \cos\varphi & & -\sin\varphi & \\ & & & \ddots & & \\ & & \sin\varphi & & -\cos\varphi & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ g \\ \\ q \\ \\ \\ \end{matrix}$$

$$\qquad\qquad g \qquad\qquad q$$

$\boldsymbol{T}_{gq}$ 中凡没有标明的元素均为 0。$\boldsymbol{A}_1$ 经过变换后，相当于将 f_g, f_q 因子平面旋转一个角度 φ，得到矩阵：

$$\boldsymbol{B} = \boldsymbol{A}_1 \boldsymbol{T}_{gq} = (b_{ij})_{m \times p}$$

$\boldsymbol{B}$ 中的元素分别为：

$$\begin{cases} b_{ig} = a_{ig} \cos\varphi + a_{iq} \sin\varphi \\ b_{iq} = -a_{ig} \cos\varphi + a_{iq} \cos\varphi \qquad (k \neq g, q; i = 1, 2, \cdots, m) \\ b_{ik} = a_{ik} \end{cases} \tag{7-10}$$

如果有 p 个主因子，则必须对 $\boldsymbol{A}_1$ 中 p 列全部配对旋转，总共需要旋转 $p(p-1)/2$ 次，全部旋转完毕算一个循环，得到载荷矩阵：

$$\boldsymbol{B}_1 = \boldsymbol{A}_1 \boldsymbol{T}_{12} \cdots \boldsymbol{T}_{1p} \cdots \boldsymbol{T}_{(p-1)p} = \boldsymbol{A}_1 \prod_{g=1}^{p-1} \prod_{q=g+1}^{p} \boldsymbol{T}_{gq} = \boldsymbol{A}_1 \boldsymbol{C}_1$$

得到 $\boldsymbol{B}_1$ 后，可按式(7-9)计算 $V_1^{\cdot}$。在第一个循环的基础上，从 $\boldsymbol{B}_1$ 出发进行第二个旋转循环，旋转完成后得到 $\boldsymbol{B}_2$，即：

$$\boldsymbol{B}_2 = \boldsymbol{B}_1 \prod_{g=1}^{p-1} \prod_{q=g+1}^{p} \boldsymbol{T}_{gq} = \boldsymbol{B}_1 \boldsymbol{C}_2 = \boldsymbol{A}_1 \boldsymbol{C}_1 \boldsymbol{C}_2$$

得到 $\boldsymbol{B}_2$ 后，即可按式(7-9)计算 $V_2^{\cdot}$。

不断重复上述计算，就可以得到 $V^{\cdot}$ 的一个非降有界序列 $V_1^{\cdot} \leqslant V_2^{\cdot} \leqslant \cdots \leqslant V_{\max}$。由于因子载荷的绝对值不大于 1，所以这个序列是有上界的，必然收敛于某一极限 $V_{\max}$（$V^{\cdot}$ 的最大值）。因此，对于所要求的计算精度 ε，当循环次数 k 充分大时，必然有：

$$|V_k^{\cdot} - V_{k+1}^{\cdot}| < \varepsilon$$

最后得到旋转后的因子载荷矩阵为：

$$\boldsymbol{B}_k = \boldsymbol{A}_1 \prod_{i=1}^{k} \boldsymbol{C}_i = \boldsymbol{A}_1 \boldsymbol{C}$$

2. 旋转角度

在任何一次旋转变换中，都涉及旋转角度的问题。确定旋转角度的步骤如下：

(1) 将式(7-10)代入式(7-9)后，将式(7-9)对 φ 求一阶导数并令其为 0，解得：

$$\tan 4\varphi = (D - 2A_1B/m)/[C - (A_1^2 - V^2)/m] = E/F \qquad (7\text{-}11)$$

式中　m——变量个数。

令

$$T_j = (a_{jg}/h_j)^2 - (a_{jq}/h_j)^2, \qquad H_j = 2(a_{jg}/h_j)(a_{jq}/h_j)$$

则

$$A_1 = \sum_{j=1}^{m} T_j, B = \sum_{j=1}^{m} H_j, C = \sum_{j=1}^{m} (T_j^2 - H_j^2), D = 2\sum_{j=1}^{m} T_j H_j$$

(2) 把式(7-9)展开，并将包含 φ 的项合并化简，最后剩下包含 $\sin 4\varphi$ 和 $\sin^2 2\varphi$ 的项，使式(7-9)成为以 $\pi/2$ 为周期的函数。因而在式(7-11)中的 4φ 只需在 $-\pi/4 \sim \pi/4$ 之间考虑即可。同时，由式(7-9)对 φ 的二阶导数应小于 0，可得：

$$E^{-1} \sin 4\varphi > 0$$

所以，φ 的符号可以根据 E 的符号确定，它应与 E 同号。因此，可按 E 和 F 的正负号来确定 4φ 所在的象限。

二、Q 型因子分析方差最大正交旋转

Q 型因子分析方差最大正交旋转与 R 型因子分析方差最大正交旋转类似，不同之处是 Q 型因子分析的主因子载荷矩阵 $\boldsymbol{A}_1 = (a_{ij})_{n \times p}$，所以，只要将以上公式中的 m 换成 n 即可。

§4　因子得分

一、R 型因子得分

主因子 f_j 是由原始变量 $x_1, x_2, \cdots, x_m$ 线性组合而成的综合变量，即

$$f_j = c_{1j}x_1 + c_{2j}x_2 + \cdots + c_{mj}x_m \qquad (j = 1, 2, \cdots, p) \qquad (7\text{-}12)$$

而因子得分是把第 i 个样品的 m 个变量的观测值代入式(7-12)计算的函数值 f_{ji}（$j=1,2,\cdots,p; i=1,2,\cdots,n$）。当 $p=2$ 时，根据 f_{1i}, f_{2i}（$i=1,2,\cdots,n$）可以在平面上对样品进行分类。

把式(7-12)改写成矩阵形式，有：

$$\boldsymbol{F} = \boldsymbol{CX} \qquad (7\text{-}13)$$

其中 $\boldsymbol{F} = (f_1, f_2, \cdots, f_p)'$，$\boldsymbol{X} = (x_1, x_2, \cdots, x_m)'$，而

$$C=\begin{bmatrix} c_{11} & c_{21} & \cdots & c_{m1} \\ c_{12} & c_{22} & \cdots & c_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ c_{1p} & c_{2p} & \cdots & c_{mp} \end{bmatrix}$$

欲求因子得分，必须先确定式(7-13)中的 $\boldsymbol{C}$。

若因子载荷矩阵 $\boldsymbol{A}_1$ 为满秩的 m 阶方阵，那么，R 型因子分析模型式(7-5)中的 $\alpha_i e_i=0$，此时有 $\boldsymbol{A}_1=\boldsymbol{A}$，故由式(7-5)可直接得 $\boldsymbol{F}=\boldsymbol{A}^{-1}\boldsymbol{X}$，即 $\boldsymbol{C}=\boldsymbol{A}^{-1}$。通常 $\boldsymbol{A}_1$ 是不满秩的 $m\times p$ 阶长方形矩阵($m>p$)，假设 $\alpha_i e_i\approx 0$，那么有 $\boldsymbol{X}\approx\boldsymbol{A}_1\boldsymbol{F}$，先对该近似式左乘 $\boldsymbol{A'}_1$ 后，再左乘 $(\boldsymbol{A'}_1\boldsymbol{A}_1)^{-1}$，则有：

$$\boldsymbol{F}=(\boldsymbol{A'}_1\boldsymbol{A}_1)^{-1}\boldsymbol{A'}_1\boldsymbol{X} \tag{7-14}$$

由式(7-13)和式(7-14)得：

$$\boldsymbol{C}=(\boldsymbol{A'}_1\boldsymbol{A}_1)^{-1}\boldsymbol{A'}_1 \tag{7-15}$$

一般说来，$\alpha_i e_i$ 将随着所选因子数的减少而变大，当它大得不可忽略时，用式(7-15)求出的系数就不能正确计算因子得分。在这种情况下，只能在最小二乘法的意义下对因子得分进行估计。为此，必须建立变量 $x_i(i=1,2,\cdots,m)$ 对因子 $f_j(j=1,2,\cdots,p)$ 的回归方程。

在因子模型已标准化的条件下，设变量 x_i 对因子 f_j 的回归方程为：

$$\hat{f}_j=b_{1j}x_1+b_{2j}x_2+\cdots+b_{mj}x_m \qquad (j=1,2,\cdots,p)$$

那么 p 个回归方程的矩阵形式为：

$$\hat{\boldsymbol{F}}=\boldsymbol{BX} \tag{7-16}$$

在式(7-16)中，$p\times n$ 阶矩阵 $\hat{\boldsymbol{F}}$ 是矩阵 $\boldsymbol{F}$ 的最小二乘解，$p\times m$ 阶矩阵 $\boldsymbol{B}$ 是回归方程系数矩阵，$\boldsymbol{X}$ 为 $m\times n$ 阶的原始数据矩阵。

分别用 $(n-1)^{-1}\boldsymbol{X'}$ 右乘式(7-16)两边，得：

$$(n-1)^{-1}\hat{\boldsymbol{F}}\boldsymbol{X'}=\boldsymbol{BXX'}(n-1)^{-1}$$

因为 $(n-1)^{-1}\boldsymbol{XX'}$ 为变量的相关矩阵 $\boldsymbol{R}$，而 $(n-1)^{-1}\hat{\boldsymbol{F}}\boldsymbol{X'}$ 是公因子与变量的相关矩阵 $\boldsymbol{A}_1$，由此得 $\boldsymbol{B}=\boldsymbol{A}_1\boldsymbol{R}^{-1}$，把 $\boldsymbol{B}$ 代入式(7-16)，得到 R 型因子得分计算式：

$$\hat{\boldsymbol{F}}=\boldsymbol{A'}_1\boldsymbol{R}^{-1}\boldsymbol{X} \tag{7-17}$$

式中 $\hat{\boldsymbol{F}}=(\hat{f}_1,\hat{f}_2,\cdots,\hat{f}_p)'$，$\boldsymbol{X}=(x_1,x_2,\cdots,x_m)'$

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix},\qquad \boldsymbol{A'}_1=\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{mp} \end{bmatrix}$$

二、Q 型因子得分

在 Q 型因子分析模型式(7-7)中，当 $\alpha_j e_j$ 不可忽略时，在最小二乘法意义下也可得到类似于式(7-17)的 Q 型因子得分计算公式。但 Q 型因子分析中诸公因子的方差收敛很快，故常用式(7-14)计算因子得分。

三、因子分析计算过程

分析 R 型和 Q 型因子分析的计算过程，给出计算机因子分析流程图(图 7-3)。

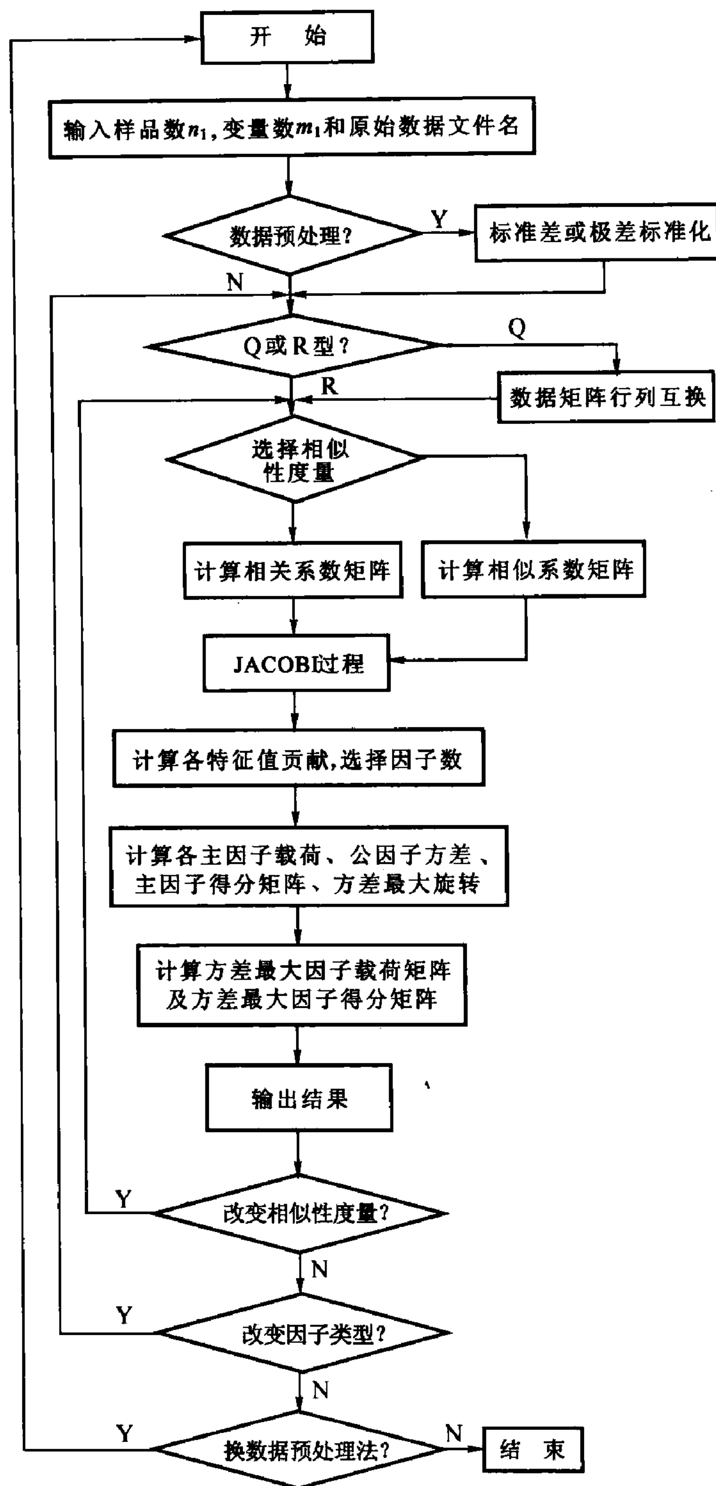

开 始
输入样品数n_1,变量数m_1和原始数据文件名
数据预处理?
Y
N
标准差或极差标准化
Q或R型?
Q
R
数据矩阵行列互换
选择相似性度量
计算相关系数矩阵
计算相似系数矩阵
JACOBI过程
计算各特征值贡献,选择因子数
计算各主因子载荷、公因子方差、主因子得分矩阵、方差最大旋转
计算方差最大因子载荷矩阵及方差最大因子得分矩阵
输出结果
改变相似性度量?
Y
N
改变因子类型?
Y
N
换数据预处理法?
Y
N
结 束

图 7-3 因子分析流程图

§5 对应分析

一、对应分析的概念

R 型和 Q 型因子分析可以用较少的几个公因子去提取研究对象的绝大部分信息，化简了研究系统，从而可在低维空间中研究样品的空间分布规律和变量成因联系，便于进行地质解释和推断。因此，因子分析在地质学中有着广泛的应用，但是，它毕竟还有不足之处：

（1）不能在同一空间研究样品和变量。

R 型和 Q 型因子分析的研究对象分别是变量和样品，这就意味着把研究样品空间分布规律和研究变量共生组合关系分隔开来。事实上，样品的特征要通过变量来揭示，如对一个既需要研究地质成因的空间分布规律，又需要研究不同类型样品特征的地质问题来说，前者就要研究样品，后者就要利用变量来解释。由此说明，地质样品和变量的研究是不可分割的，也就是说应该找出一种把两种因子分析统一起来的研究方法。

（2）占机时和内存多。

样品的数量一般远比变量的数目多，计算样品相似系数时不仅浪费机时，而且占用大量的内存空间。

（3）地质数据的尺度不同。

为了使地质数据在同一尺度下参与地质分析，往往是将变量进行标准化处理，然而对样品就不好进行标准化了。对变量和样品的非对等标准化处理，会造成地质数据的尺度不同，从而影响对地质规律的认识。

鉴于上述原因，在以上两种因子分析的基础上产生了对应分析。它采用一种数据处理方法，把两种因子分析结合起来，由 R 型因子分析结果直接导出 Q 型因子分析结果，并在同一空间内对变量和样品进行分类，由此既可研究样品的分布规律，又可通过变量对样品进行地质解释。

二、对应分析的数据变换

假设有 n 个样品，每个样品有 m 个变量，它们的观测值为：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

其中 $x_{ij}\geqslant 0(i=1,2,\cdots,m;j=1,2,\cdots,n)$，并且在每一行和每一列上至少有一个不为 0。

1. 原始数据总和标准化

数据矩阵 $\boldsymbol{X}$ 中所有元素之和为：

$$T=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}$$

数据矩阵 $\boldsymbol{X}$ 中的每个元素除以 $\boldsymbol{T}$，得总和标准化数据矩阵：

$$\boldsymbol{P}=(p_{ij})_{m\times n}$$

2. 样品坐标的相对比例

用矩阵 $\boldsymbol{P}$ 中第 j 列上元素之和

$$p_{.j}=\sum_{i=1}^{m}p_{ij}\qquad(j=1,2,\cdots,n)$$

去除 $\boldsymbol{P}$ 中第 j 列上的每个元素，得：

$$(p_{1j}/p_{.j},p_{2j}/p_{.j},\cdots,p_{mj}/p_{.j})'\quad(j=1,2,\cdots,n)$$

上式表示 m 维空间中的 n 个样品点，每个样品点的坐标是各个变量在该样品中的相对比例。因此，研究样品的相似性就变为研究 m 维空间中样品点的相对位置。样品点间的距离越小，样品间的性质就越相似。

3. 样品点间的加权距离

计算样品点 l 与 k 之间的距离时，考虑到指标数量级的差异，采用加权距离

$$\begin{aligned}D(l,k)&=\sqrt{\sum_{i=1}^{m}(p_{il}/p_{.l}-p_{ik}/p_{.k})^{2}/p_{i.}}\\&=\sqrt{\sum_{i=1}^{m}\left[p_{il}/(p_{.l}\sqrt{p_{i.}})-p_{ik}/(p_{.k}\sqrt{p_{i.}})\right]^{2}}\end{aligned}\quad(7\text{-}18)$$

式中 $p_{i.}$——$\boldsymbol{P}$ 中第 i 行上元素之和，即：

$$p_{i.}=\sum_{j=1}^{n}p_{ij}\qquad(i=1,2,\cdots,m)$$

它是第 i 个变量的权。由式(7-18)可知，计算加权距离 $D(l,k)$ 时，需要把 n 个样品点的坐标改为：

$$(p_{1j}/(p_{.j}\sqrt{p_{1.}}),p_{2j}/(p_{.j}\sqrt{p_{2.}}),\cdots,p_{mj}/(p_{.j}\sqrt{p_{m.}}))'\qquad(j=1,2,\cdots,m)$$

4. 变量坐标的相对比例

用 $\boldsymbol{P}$ 中第 i 行上的元素之和

$$p_{i.}=\sum_{j=1}^{n}p_{ij}\qquad(i=1,2,\cdots,m)$$

去除 $\boldsymbol{P}$ 中第 i 行上的每个元素，得：

$$(p_{i1}/p_{i.},p_{i2}/p_{i.},\cdots,p_{in}/p_{i.})\qquad(i=1,2,\cdots,m)$$

上式表示 n 维空间中的 m 个变量点。

5. 变量点间的加权距离

任意两点 l 与 k 之间的加权距离为：

$$\begin{aligned}D^{*}(l,k)&=\sqrt{\sum_{j=1}^{n}(p_{lj}/p_{l.}-p_{kj}/p_{k.})^{2}/p_{.j}}\\&=\sqrt{\sum_{j=1}^{n}\left[p_{lj}/(p_{l.}\sqrt{p_{.j}})-p_{kj}/(p_{k.}\sqrt{p_{.j}})\right]^{2}}\end{aligned}\quad(7\text{-}19)$$

由式(7-19)可知，为了计算加权距离 $D^{*}(l,k)$，需要把 m 个变量点的坐标改为：

$$(p_{i1}/(p_{i.}\sqrt{p_{.1}}),p_{i2}/(p_{i.}\sqrt{p_{.2}}),\cdots,p_{in}/(p_{i.}\sqrt{p_{.n}}))\qquad(i=1,2,\cdots,m)$$

三、协方差矩阵

1. 变量的协方差矩阵

(1) 变量的均值。

若将矩阵 $\boldsymbol{P}$ 中的元素 p_{ij} 视为概率，那么 $p_{i.}$，$p_{.j}$ 就是边缘概率。因此，m 维空间中样品点第 i 个变量的概率均值为：

$$\sum_{j=1}^{n} p_{ij}/(\sqrt{p_{i.}}\,p_{.j}) \cdot p_{.j} = \sqrt{p_{i.}} \quad (i=1,2,\cdots,m)$$

(2) 变量的协方差矩阵。

第 i 个与第 j 个变量的协方差为：

$$\begin{aligned} s_{ij} &= \sum_{k=1}^{n}[p_{ik}/(\sqrt{p_{i.}}\,p_{.k}) - \sqrt{p_{i.}}][p_{jk}/(\sqrt{p_{j.}}\,p_{.k}) - \sqrt{p_{j.}}] \cdot p_{.k} \\ &= \sum_{k=1}^{n}[(p_{ik} - p_{i.}\,p_{.k})/\sqrt{p_{i.}\,p_{.k}}][(p_{jk} - p_{j.}\,p_{.k})/\sqrt{p_{j.}\,p_{.k}}] \\ &= \sum_{k=1}^{n} z_{ik} \cdot z_{jk} \end{aligned}$$

式中

$$\begin{aligned} z_{ik} &= (p_{ik} - p_{i.}\,p_{.k})/\sqrt{p_{i.}\,p_{.k}} \\ &= [x_{ik}/T - (x_{i.}/T) \cdot (x_{.k}/T)]/[(x_{i.}/T) \cdot (x_{.k}/T)]^{2} \\ &= (x_{ik} - x_{i.}\,x_{.k}/T)(x_{i.}\,x_{.k})^{-1/2} \end{aligned}$$

$$x_{i.} = \sum_{k=1}^{n} x_{ik} \quad (i=1,2,\cdots,m)$$

$$x_{.k} = \sum_{k=1}^{m} x_{ik} \quad (i=1,2,\cdots,n)$$

按上式计算，m 个变量的协方差矩阵为：

$$\boldsymbol{S} = (s_{ij})_{m\times m} = \boldsymbol{ZZ}'$$

式中

$$\boldsymbol{Z} = (z_{ik})_{m\times n}$$

2. 样品的协方差矩阵

(1) 样品的均值。

在 n 维空间中，第 k 个样品的概率均值为：

$$\sum_{i=1}^{m} p_{ik}/(p_{i.}\ \sqrt{p_{.k}}) \cdot p_{i.} = \sqrt{p_{.k}} \quad (k=1,2,\cdots,n)$$

(2) 样品的协方差矩阵。

任意两个样品 l 和 k 的协方差为：

$$\begin{aligned} s_{kl}^{*} &= \sum_{i=1}^{m}[p_{ik}/(p_{i.}\ \sqrt{p_{.k}}) - \sqrt{p_{.k}}][p_{il}/(p_{i.}\ \sqrt{p_{.l}}) - \sqrt{p_{.l}}] \cdot p_{i.} \\ &= \sum_{i=1}^{m}[(p_{ik} - p_{i.}\,p_{.k})/\sqrt{p_{i.}\,p_{.k}}][(p_{il} - p_{i.}\,p_{.l})/\sqrt{p_{i.}\,p_{.l}}] \\ &= \sum_{i=1}^{m} z_{ik} \cdot z_{il} \end{aligned}$$

式中

$$z_{ik} = (x_{ik} - x_{i.}\,x_{.k}/T)(x_{i.}\,x_{.k})^{-1/2}$$

按上式计算，n 个样品的协方差矩阵为：

$$\boldsymbol{S}^{*} = (s_{kl})_{n\times n} = \boldsymbol{Z}'\boldsymbol{Z}$$

式中

$$Z = (z_{ik})_{n\times n}$$

四、因子载荷矩阵

由线性代数可知，矩阵 ZZ' 与 $Z'Z$ 有相同的非零特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p (p \leqslant m)$，并且对于每个 $\lambda_j (1 \leqslant j \leqslant p)$，若对应的 u_j 是 ZZ' 的单位特征向量，那么 $v_j = Z'u_j$ 是 $Z'Z$ 相应的单位特征向量；反之，若 v_j 是与 $Z'Z$ 的特征值 λ_j 对应的单位特征向量，那么 $u_j = Zv_j$ 则是 ZZ' 所相应的单位特征向量。

上述结果表明，当求得变量的协方差矩阵 S 的特征值 $\lambda_j (1 \leqslant j \leqslant p)$ 和与其对应的特征向量 $u_j (j = 1,2,\cdots,p)$ 后，便可得到 R 型因子分析的因子载荷矩阵，再由 $v_j = Z'u_j$ 直接求得 Q 型因子分析的因子载荷矩阵。此外，S 与 S^* 有相同的特征值，这些特征值表示各公因子所提供的方差。因此变量空间中的 p 个因子与样品空间中的 p 个因子在总方差中所占的百分比完全相同，故用同样的因子轴既可以表示变量，又可以表示样品，从而把 R 型和 Q 型因子分析统一起来。

1. R 型因子载荷矩阵

若取 S 的前 p 个特征值 $\lambda_j (1 \leqslant j \leqslant p)$，与之对应的特征向量为 $u_j (j = 1,2,\cdots,p)$，那么 R 型因子载荷矩阵为：

$$U = \begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1p}\sqrt{\lambda_p} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2p}\sqrt{\lambda_p} \\ \vdots & \vdots & \cdots & \vdots \\ u_{m1}\sqrt{\lambda_1} & u_{m2}\sqrt{\lambda_2} & \cdots & u_{mp}\sqrt{\lambda_p} \end{pmatrix}$$

2. Q 型因子载荷矩阵

根据 $v_j = Z'u_j$，Q 型因子载荷矩阵为：

$$V = \begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1p}\sqrt{\lambda_p} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2p}\sqrt{\lambda_p} \\ \vdots & \vdots & \cdots & \vdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \cdots & v_{np}\sqrt{\lambda_p} \end{pmatrix}$$

五、对应分析计算步骤

1. 求 Z 矩阵

把原始数据矩阵 X 变换为 Z 矩阵：

$$z_{ij} = (x_{ij} - x_{i.}x_{.j}/T)\Big/\sqrt{x_{i.}x_{.j}} \qquad (i = 1,2,\cdots,m; j = 1,2,\cdots,n)$$

式中

$$x_{i.} = \sum_{k=1}^{n} x_{ik} \qquad (i = 1,2,\cdots,m)$$

$$x_{.j} = \sum_{k=1}^{m} x_{kj} \qquad (j = 1,2,\cdots,n)$$

$$T = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$$

2. R 型因子分析

求变量协方差矩阵 S 的特征值 $\lambda_j (j = 1,2,\cdots,m)$，按 $\left(\sum_{i=1}^{p}\lambda_j \Big/ \sum_{j=1}^{m}\lambda_j\right)$ 达到的要求取前 p

个主因子，计算因子载荷矩阵 $\boldsymbol{U}$。

3. Q 型因子分析

根据 $\boldsymbol{v}_j=\boldsymbol{Z}'\boldsymbol{u}_j$ 计算 Q 型因子载荷矩阵 $\boldsymbol{V}$。

4. 作图及地质解释

在因子平面内，以因子载荷为坐标做变量和样品散点图，进行地质解释。

§6 应用实例

【例 1】砂岩分类。

根据砂岩的主要碎屑(石英、长石与岩屑)的百分含量对砂岩分类。现有东濮凹陷 18 块砂岩样品分析数据(表 7-1，据赵旭东，修改)，这些样品均有人工鉴定结果，重新进行对应分析的目的在于检验统计分析方法的有效性。两个特征值累计百分比为 100%。在对应分析平面图(图 7-4)上，样品分为四类，与人工鉴定分类命名结果完全吻合。该例表明，对应分析是进行分类和成因解释的一种有效方法。

表 7-1 岩样分析数据及岩石名称

样品号	石英含量/%	长石含量/%	岩屑含量/%	其他含量/%	鉴定命名
1	82	8	9	1	石英砂岩
2	83	8	9	0	石英砂岩
3	86	7	5	2	石英砂岩
4	88	8	12	2	硬砂质石英砂岩
5	78	7	12	3	硬砂质石英砂岩
6	84	5	11	0	硬砂质石英砂岩
7	76	9	13	2	硬砂质石英砂岩
8	61	26	13	0	长石砂岩
9	60	25	15	0	长石砂岩
10	58	29	13	0	长石砂岩
11	57	27	16	0	长石砂岩
12	55	30	15	0	长石砂岩
13	47	27	26	0	混合砂岩
14	46	25	29	0	混合砂岩
15	58	7	35	0	硬砂岩
16	53	17	30	0	长石质硬砂岩
17	61	12	25	2	长石质硬砂岩
18	50	17	33	0	长石质硬砂岩

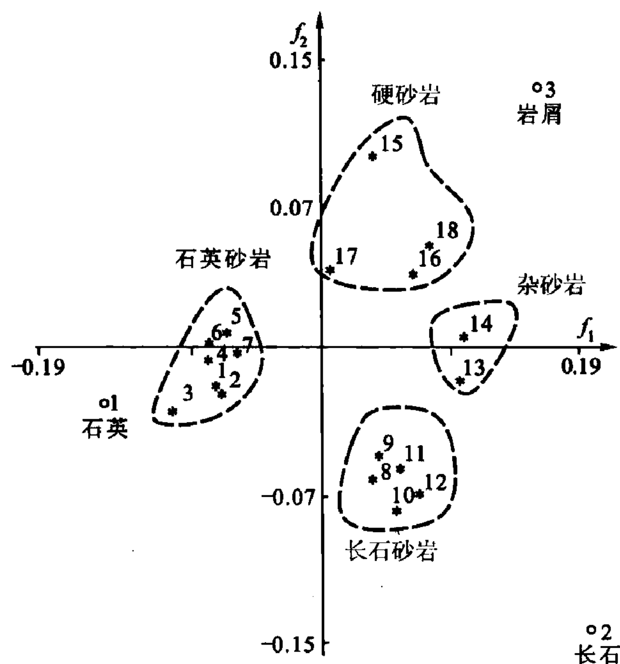


图 7-4 砂岩样品与碎屑对应分析平面图

【例 2】古潜山油气藏与成油地质参数分析。

对已探明的一些古潜山油气藏进行分类，研究它们的资源量与成油地质参数的关系，是评价未知古潜山油气资源量的基础。与其他类型的油气藏一样，油源、供油、储集、圈闭和保存条件等是形成古潜山油气藏的基本条件，将其拟定为如下八项成油地质参数：

x_1——古潜山到凹陷生油区中心的距离，km；

x_2——储集体的总孔隙度，%；

x_3——古潜山上覆盖层的厚度，km；

x_4——供油窗口，km^2；

x_5——生油岩与不整合面的接触面积,km^2;

x_6——圈闭闭合高度,km;

x_7——圈闭闭合面积,km^2;

x_8——古潜山最浅埋藏深度,km。

由于存在同一个生油凹陷向不同古潜山圈闭提供油气的问题,本例未把油源条件列入成油地质参数。对已探明的26个古潜山油气藏及其成油地质参数(表7-2)进行对应分析,计算的有关统计参数见表7-3、表7-4。

表7-2 古潜山油气藏成油地质参数

油藏序号	成油地质参数							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	4.0	3.17	1.26	0.72	13.5	0.70	13.0	2.40
2	12.0	4.67	1.30	0.50	6.50	0.45	6.00	2.30
3	12.0	8.60	0.50	2.60	6.00	0.30	5.20	0.11
4	8.0	10.00	0.80	0.02	4.00	0.30	3.20	1.40
5	8.0	4.14	1.00	0.02	4.00	0.35	4.00	1.70
6	2.0	3.99	2.35	6.40	10.60	1.20	27.00	2.60
7	3.0	3.86	2.23	0.30	9.40	0.50	9.70	2.40
8	3.0	3.32	1.60	1.44	11.70	0.70	8.40	1.60
9	5.5	3.40	0.56	1.64	5.32	0.35	5.90	0.65
10	3.5	3.70	2.28	5.00	26.60	0.70	26.60	2.50
11	17.5	3.72	2.60	16.00	183.00	1.91	183.00	2.60
12	6.5	2.79	3.97	3.00	4.30	0.36	4.30	3.97
13	2.5	14.03	2.30	10.00	24.00	1.00	24.00	2.30
14	9.0	2.79	2.80	10.00	1.20	0.25	1.20	2.80
15	4.0	2.80	5.05	6.25	4.00	0.45	4.00	5.05
16	3.0	4.16	4.04	1.29	5.50	0.26	5.50	4.04
17	4.0	2.76	4.50	6.75	4.00	0.30	4.00	4.50
18	5.0	4.71	3.85	2.00	6.24	0.55	6.24	3.85
19	8.0	5.82	3.10	9.38	1.84	0.20	1.84	3.10
20	10.0	3.17	1.95	7.50	5.90	0.40	5.90	1.95
21	8.0	1.93	3.15	12.00	3.40	0.35	3.40	3.15
22	20.0	2.50	3.75	6.00	7.00	0.60	7.00	3.75
23	20.0	1.74	4.40	6.00	16.54	0.95	16.50	4.00
24	20.0	1.42	3.70	6.00	4.00	0.65	4.00	3.70
25	8.0	5.50	0.95	5.80	70.00	1.00	64.90	3.50
26	1.0	18.00	0.93	0.15	0.20	0.60	1.80	2.15

表7-3 特征值及R型因子分析第1,2公因子载荷

特征值序号	特征值	特征值百分数/%	特征值累积百分数/%	变量序号	因子载荷	
					f_1	f_2
1	0.319 5	57.91	57.91	1	0.261 7	−0.106 7
2	0.129 5	23.47	81.38	2	0.215 5	0.308 4
3	0.063 8	11.56	92.94	3	0.156 1	−0.060 3
4	0.030 9	5.60	98.54	4	0.163 7	−0.135 2
5	0.006 2	1.12	99.66	5	−0.256 5	0.002 6
6				6	0.027 2	0.014 6
7				7	−0.252 7	0.009 8
8				8	0.151 8	−0.028 0

表 7-4　Q 型因子分析第 1,2 公因子载荷

油藏序号	公因子载荷 f_1	公因子载荷 f_2	油藏序号	公因子载荷 f_1	公因子载荷 f_2
1	−0.029 9	0.018 6	14	0.138 1	−0.067 6
2	0.070 6	0.010 3	15	0.098 0	−0.038 7
3	0.088 6	0.054 3	16	0.057 7	0.019 5
4	0.098 1	0.109 4	17	0.094 4	−0.040 0
5	0.066 0	0.025 8	18	0.077 3	0.004 8
6	−0.034 6	0.003 6	19	0.142 6	−0.021 1
7	−0.002 8	0.031 2	20	0.082 6	−0.042 6
8	−0.012 7	0.020 8	21	0.117 1	−0.084 5
9	0.028 4	0.018 0	22	0.117 2	−0.072 3
10	−0.072 1	0.002 8	23	0.051 9	−0.067 9
11	−0.323 0	−0.018 2	24	0.139 9	−0.090 9
12	0.085 4	−0.023 9	25	−0.169 5	0.010 1
13	−0.005 7	0.074 2	26	0.142 6	0.263 8

古潜山油气藏分类结果见对应分析第 1,2 公因子平面图(图 7-5)。两个公因子能够提取原始数据信息量的 81.38%,并且第 1 个公因子占了 57.91%,由此可见,它是形成古潜山油气藏的主导性地质因素。成油地质参数沿 f_1 轴从左向右的分布顺序依次为 $x_5,x_7,x_6,x_3,x_8,x_4,x_2,x_1$。根据成油地质参数的分布及其意义,可以将 f_1 因子理解为油气的地质输导作用。根据成油地质参数沿 f_2 轴的分布特征,可视 f_2 的地质意义为构造运动和风化剥蚀作用。

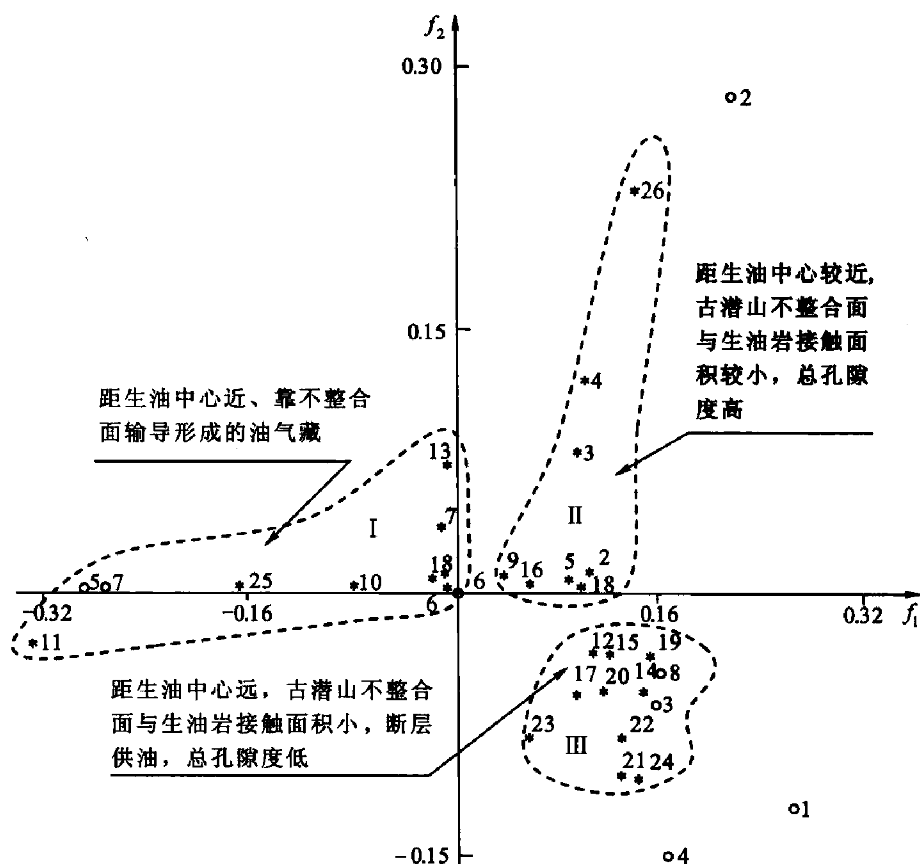


图 7-5　油气藏与成油地质参数对应分析平面图

在对应分析平面图上,26 个古潜山油气藏分为三类:

Ⅰ类古潜山油气藏距生油凹陷中心近,古潜山不整合面与生油岩接触面积大,闭合面积和闭合高度都较大,典型的特征是输导条件好。该类油气藏的平均探明储量在 100×10^6 t 以上,个别油藏超过 500×10^6 t。

Ⅱ类古潜山油气藏距生油凹陷中心较近,古潜山不整合面与生油岩接触面积较小,总孔隙度高。该类油气藏的平均探明储量为 10×10^6 t。

Ⅲ类古潜山油气藏距生油凹陷中心远,古潜山不整合面与生油岩接触面积小,靠断层供油,总孔隙度低。该类油气藏的平均探明储量为 5×10^6 t。

综上所述,近油源的不整合面与生油岩接触面积较大的褶皱山和残山容易形成含油气丰富的古潜山油气藏。

【例 3】盐泉分类与成因。

为了对云南某地盐泉进行分类和成因解释,对云南某地 20 个盐泉水化学分析数据(表 7-5)进行对应分析。

表 7-5 盐泉水化学分析数据

盐泉序号	矿化度 /(g·L^{-1})	$10^3\ Br/Cl$	$10^3\ K/\sum$ 盐	$10^3\ K/Cl$	Na/K	$10^2\ Mg/Cl$	$\varepsilon(Na)/\varepsilon(Cl)$
851	11.853	0.480	14.360	25.210	25.210	0.810	0.980
2	45.596	0.526	13.850	24.040	26.010	0.910	0.960
3	3.525	0.086	24.400	49.300	11.300	6.820	0.850
4	3.681	0.370	13.570	25.120	26.000	0.820	1.010
5	48.28	0.386	14.500	25.900	23.320	2.180	0.930
6	17.956	0.280	9.750	17.050	37.200	0.464	0.980
7	7.370	0.506	13.600	34.210	10.690	8.800	0.560
8	4.223	0.340	3.800	7.100	88.200	1.110	0.970
9	6.442	0.190	4.700	9.100	23.200	0.740	1.080
10	16.234	0.390	3.400	5.400	121.500	0.420	1.000
11	10.585	0.420	2.400	4.700	135.600	0.870	0.980
12	23.535	0.230	2.600	4.600	141.800	0.310	1.020
13	5.398	0.120	2.800	6.200	111.200	1.140	1.070
14	283.148	0.148	1.763	2.968	215.860	0.140	0.980
15	316.604	0.317	1.453	2.432	263.410	0.249	0.980
16	307.310	0.173	1.627	2.729	235.700	0.214	0.990
17	322.515	0.312	1.382	2.320	282.210	0.024	1.000
18	256.580	0.297	0.899	1.476	410.300	0.239	0.930
19	304.092	0.283	0.789	1.357	438.360	0.193	1.010
20	240.446	0.042	0.741	1.266	500.770	0.290	0.990

矩阵 $\boldsymbol{ZZ'}$ 的前两个特征值所代表的方差已占总方差的 95.94%。因此,取前两个主因子就可很好地反映原始数据的变化。前两个主因子载荷见表 7-6。

由于第一个主因子 f_1 的方差占总方差的 79%以上,因此可以说,它是区内起主导作用的一个因素,基本上能够反映本区沉积环境演变的主要特征。在对应分析因子平面图(图 7-6)上,f_1 轴的左端为钠,右端为钾,含钾盐泉的主要特征变量 $10^3\ K/Cl$, $10^3\ K/\sum$ 盐,

$10^2\ Mg/Cl$,$\varepsilon(Na)/\varepsilon(Cl)$,$10^3\ Br/Cl$ 都分布在 f_1 轴的右端，严格受 f_1 的控制。在 f_1 因子载荷中，占比重最大的变量是 $10^3\ K/Cl$，位于 f_1 轴的最右端。对于其他变量，按其因子载荷在 f_1 中所占比例大小自右至左依次为 $10^3\ K/\sum$ 盐，$10^2\ Mg/Cl$,$\varepsilon(Na)/\varepsilon(Cl)$,$10^3\ Br/Cl$,$Na/K$，从而反映出各种盐类物质随着沉积环境的变化而开始沉积的先后次序。钾的浓度自左至右递增，而钠的浓度变化却相反。因此，可以把 f_1 视为盐类沉积分异作用。钠位于 f_1 轴的左端，反映了富钠的沉积环境。$10^3\ Br/Cl$,$10^2\ Mg/Cl$ 靠近坐标原点，说明镁盐与溴化物在本区钠盐和钾盐的沉积过程中具有一定的浓度，并且钾盐沉积过程中镁盐的混入要比溴化物更为明显。因子 f_2 提供的方差要比 f_1 小得多，仅占总方差的 16.30%。因子载荷较大的变量是矿化度和 Na/K，分别位于 f_2 轴的上、下端，这两个变量主要受 f_2 因素支配，而其他变量分布在 f_2 轴的中部，在 f_2 上的因子载荷都很小。该因子表明富钠环境的沉积阶段，杂质的混入更为明显。

表 7-6　R 型与 Q 型前两个因子载荷

变量序号	变量名	f_1	f_2	样品序号	f_1	f_2	样品序号	f_1	f_2
1	矿化度/(g·L^{-1})	−0.144 1	0.228 5	1	0.200 0	−0.000 7	11	−0.005 1	−0.112 9
2	$10^3\ Br/Cl$	0.034 2	−0.008 3	2	0.146 4	0.048 8	12	−0.005 1	−0.112 9
3	$10^3\ K/\sum$ 盐	0.347 2	0.010 4	3	0.384 8	0.014 2	13	−0.005 1	−0.112 9
4	$10^3\ K/Cl$	0.504 1	0.014 4	4	0.210 8	−0.017 6	14	−0.076 0	0.088 7
5	Na/K	−0.115 9	−0.197 7	5	0.161 0	0.056 0	15	−0.076 0	0.088 7
6	$10^2\ Mg/Cl$	0.195 2	0.002 3	6	0.119 7	−0.009 3	16	−0.080 7	0.091 6
7	$\varepsilon(Na)/\varepsilon(Cl)$	0.045 4	−0.020 9	7	0.307 6	0.015 3	17	−0.086 7	0.074 6
				8	0.029 1	−0.091 9	18	−0.088 4	−0.027 1
特征值		0.450 3	0.092 2	9	0.089 4	−0.021 4	19	−0.095 3	−0.010 0
特征值累积百分比/%		79.65	95.94	10	0.000 5	−0.094 3	20	−0.092 1	−0.073 7

依据上述分析，按变量与样品的分布情况，可把样品分为三个区：Ⅰ区位于 f_1 轴右端，样品特征偏于含钾，表明样品区有利于形成钾盐矿床；Ⅱ区位于 f_2 轴左侧，样品特征偏于含钠，在 f_1 轴上的因子载荷很相近，而在 f_2 轴上的因子载荷差别较大，表明矿化度对钠盐有较大的影响；Ⅲ区位于上述两区之间，主要受 f_2 的影响，在靠近 Na/K 的方向，盐泉有一定程度的钾矿化，沉积顺序为过渡型，盐泉中有一定程度的钠，另外，其他混入物质也相对增加。

【例 4】大王北洼陷油源因子分析。

根据第四章【例 3】资料对大王北洼陷油样和烃源岩样品作因子分析，进行油源对比。前两个主因子 f_1,f_2 包含了原始变量的绝大部分信息。从因子分析图(图 7-7)看到：DB25-23，DB10-4，D371，D65-51，D65 原油样品和沙三段烃源岩样品集中分布于 f_1 右端，因子载荷较大，在因子 f_2 上的载荷为负值，表明这些原油样品和沙三段烃源岩有密切成因联系。DX361 和 D359-2 原油样品与沙四段烃源岩样品分布于 f_1 右端的上下侧，因子载荷较高，表明这两个原油样品与沙四段烃源岩具有明显的成因联系。D359-1 原油样品和沙一段烃源岩样品在 f_1 和 f_2 上的因子载荷相差不大，表明他们之间有密切的成因联系。在一定范围内，样品 f_2 载荷的大小与样品的成熟度参数具有负相关性，当成熟度达到某个临界值时，样品在 f_1（成熟度）上的载荷趋于稳定。因此，根据 D359-1，D35-5-4，D35-11-x5，D355，DB14-18 原油样品的分布趋势，上述原油样品应与 Es_1 烃源岩有成因联系，但也可能是多源

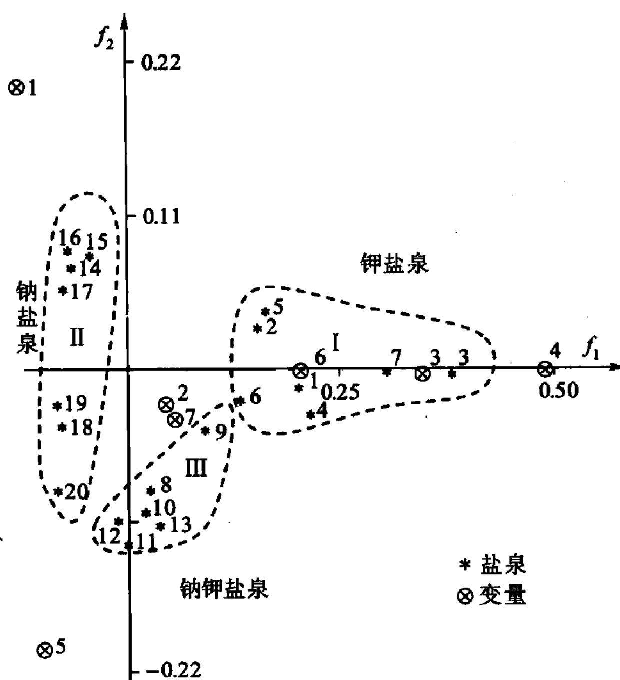
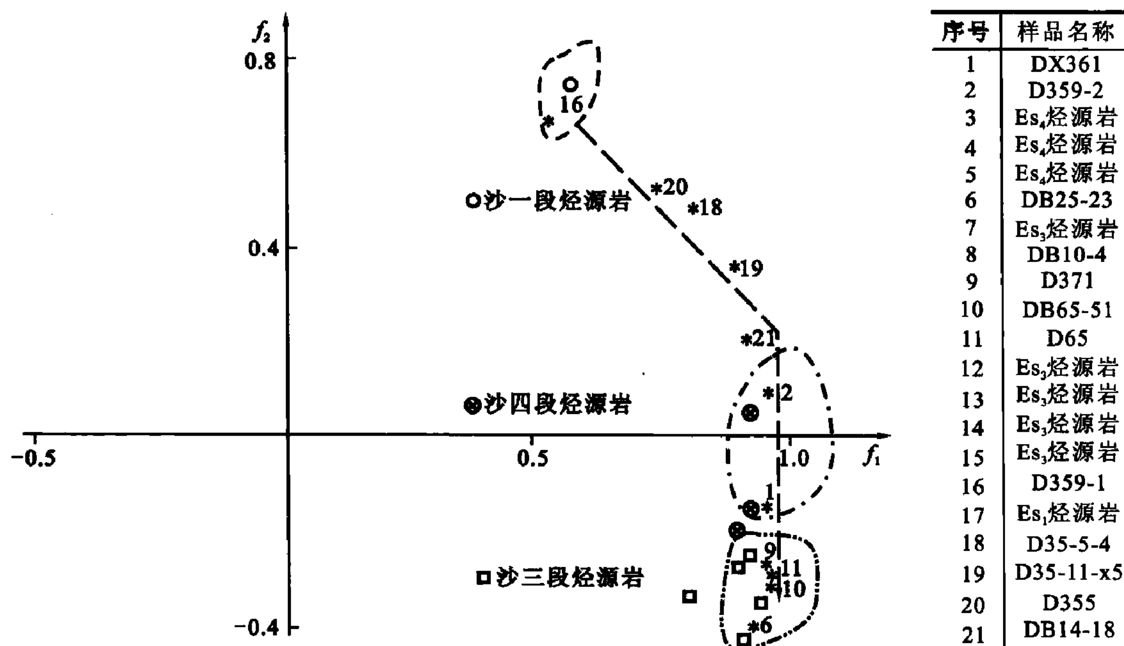


图 7-6 盐泉与水化学对应分析图

混合成因。

其余原油样品和各层段烃源岩都没有密切的相关性，但均分布于各烃源岩的过渡带上，可能表明他们是多源混合成因的。所有样品在 f_2 上载荷较大，载荷的大小与样品的成熟度参数具有负相关关系，因子 f_2 在某种意义上反映了沉积环境的差异。



序号	样品名称
1	DX361
2	D359-2
3	Es_4烃源岩
4	Es_4烃源岩
5	Es_4烃源岩
6	DB25-23
7	Es_3烃源岩
8	DB10-4
9	D371
10	DB65-51
11	D65
12	Es_3烃源岩
13	Es_3烃源岩
14	Es_3烃源岩
15	Es_3烃源岩
16	D359-1
17	Es_1烃源岩
18	D35-5-4
19	D35-11-x5
20	D355
21	DB14-18

图 7-7 原油样品与烃源岩样品因子分析平面图(据任拥军，2006)

思考与练习

1. 什么是因子分析？它的主要功能是什么？
2. 何谓R型和Q型因子分析，它们的研究对象是什么？
3. 因子分析可用来研究地学中哪些方面的问题？
4. R型和Q型因子分析有何不同？
5. 熟悉因子分析模型，并理解模型的意义。
6. 如何求因子分析中的主因子解？
7. 在因子分析中，为何要进行方差最大正交旋转？
8. 什么是因子得分？如何计算因子得分？
9. 试述因子分析的计算步骤，如何应用得到的各种结果？
10. 什么是对应分析？在对应分析图上可以获得哪些信息？

第八章　蒙特卡罗法

§1　蒙特卡罗法概述

一、蒙特卡罗法的基本思想

蒙特卡罗法又叫做统计试验法。它是针对数值解不确定的一类问题，求其在某条件下的概率解的一种统计学方法。其基本思想可概括为：欲求某问题的概率解，就构造一个表征该问题概率解的数学模型，记为

$$Y = \mu(X_1, X_2, \cdots, X_n) \tag{8-1}$$

使问题的概率解是式(8-1)的某个数字特征(如数学期望、方差等)，并且这个数字特征又能够用统计的方法求得其估计值，那么就把该估计值作为该问题概率解的近似值。

二、求解的基本过程

由蒙特卡罗法的基本思想可以得出，求解过程大致可归纳为四步：

(1) 分析数值解不确定问题所依赖的随机变量，构造表征数值解不确定问题概率解的数学模型。

(2) 对概率解数学模型式(8-1)中的随机变量 $X_1, X_2, \cdots, X_n$进行 m 次随机抽样，获得随机变量的 m 组抽样值：

$$x_{1k}, x_{2k}, \cdots, x_{nk} \quad (k = 1, 2, \cdots, m) \tag{8-2}$$

(3) 分别把式(8-2)中的 m 组抽样值代入式(8-1)，求出随机变量 Y 的 m 个估计值 $y_1, y_2, \cdots, y_m$。

(4) 根据 $y_1, y_2, \cdots, y_m$，用频率统计法求出 Y 的分布曲线(图 8-1)，由此可以获得概率不小于 p_i 所对应的解 y_{p_i}。

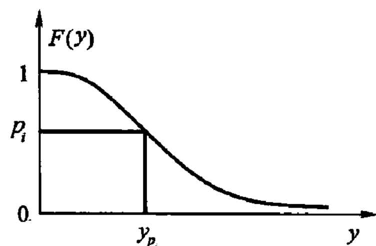


图 8-1　分布曲线与概率解示意图

三、方法中的核心问题

由求解过程可知，利用蒙特卡罗法求数值解不确定问题概率解的核心是对随机变量进行随机抽样，获得其抽样值。但是，对于数学、物理、工程技术、地质学等问题中的随机变量往往不可能进行 m 次随机抽样，解决该问题的方法是先用数学的方法在计算机上产生数以千计、万计，甚至是百万计的在[0,1]区间上均匀分布的随机数，然后利用这些随机数来实现对不同分布的随机变量进行随机抽样。

综上所述，蒙特卡罗法是以概率论与数理统计理论为指导，有着广泛应用领域的通用性统计学方法。20 世纪 40 年代，在计算机上对中子的行为进行随机抽样模拟，推断所要求的参数。1946 年，物理学家冯・诺依曼(John von Neumann)用随机抽样的方法模拟了中子连锁反应。当时出于保密，将这种统计学方法以赌城蒙特卡罗的名字命名为蒙特卡罗法。

油气资源评价属于数值解不确定的问题，20 世纪 60 年代开始把蒙特卡罗法用于油气资源的定量估算。美国 1975 年完成的第二次全美石油资源评价的主要算法就是蒙特卡罗法。目前世界上各主要产油国及西方各大石油公司都把这种方法作为石油资源定量评价的

重要方法之一，广泛应用于含油气区的早、中期勘探阶段。我国应用该方法估算石油资源量始于 1975 年，是第二次全国油气资源评价的主要算法之一，现在国内各石油公司已普遍使用，并已成为以统计预测为主的应用软件评价系统的核心算法。

§2 随机数的产生和检验

随机数是随机变量的观测值，由其构成的数据序列叫做随机数序列，它是一个无周期的数据序列。如前所述，要对数值解不确定问题中的随机变量进行随机抽样，则需要先用数学的方法在计算机上产生数以千计、万计，甚至是百万计的在[0,1]区间上均匀分布的随机数，这种随机数构成一个周期性的数据序列。为此，称这种周期性数据序列为“伪随机数”序列，其中的元素叫做“伪随机数”。在伪随机数制约下所获得的随机变量的抽样值序列显然不是一个随机序列，也就是说，这种抽样值并非是随机变量的真实观测值。尽管如此，只要对伪随机数序列进行一系列严格的统计检验，证明它可以满足模拟计算的精度要求，则伪随机数就可以作为真随机数使用。

为了满足模拟问题的需要，在计算机上产生的伪随机数序列不仅要有足够长的周期，而且应当具有符合要求的概率统计性质。

从理论上讲，只要有一种连续分布的随机数，就可以采用数学变换的方法产生其他分布的随机数。[0,1]区间上均匀分布的随机变量的抽样值是最简单、最基本的一种连续分布的随机数，其他分布的随机数都可以借助它来产生，所以说均匀分布随机数是实现随机抽样的基本工具。

一、随机数的产生方法

在计算机上产生伪随机数的数学方法有迭代取中法、移位法和同余法等。前两种方法所产生的伪随机数序列的周期对初始值的依赖性很大，选得不好时，伪随机数序列的长度较短，不能满足模拟计算的需要。目前产生伪随机数序列比较好的方法是同余法中的乘同余法和混合同余法。

1. 乘同余法

该方法产生伪随机数序列的递推同余公式为：

$$\begin{cases} x_{n+1} \equiv \alpha x_n \qquad (\mathrm{mod}\ M) \\ r_{n+1} = x_{n+1}/M \end{cases} \tag{8-3}$$

式中 x_n, x_{n+1}——分别是第 n 次和第 $n+1$ 次产生的伪随机数；

α——乘子系数；

M——模；

r_{n+1}——[0,1]区间上的第 $n+1$ 个伪随机数。

$x_{n+1} \equiv \alpha x_n (\mathrm{mod}\ M)$ 叫做以 M 为模的同余式，其含义为 x_{n+1} 是 α 与 x_n 的乘积除以 M 的余数部分。

乘同余法由 Lehmer 提出，他曾取 $M=100\ 000\ 001$，$\alpha=23$，初值 $x_0=47\ 594\ 118$，得到周期为 5 882 352 的 8 位十进制伪随机数序列，并对其中的 5 000 个伪随机数进行统计检验，其结果认为是满意的。

乘同余法所产生的伪随机数序列的周期与初值 x_0、乘子系数 α 有密切的关系。因此，用乘同余法产生伪随机数时，设计算机字长为 k，$M=2^k$，取 x_0 与 M 互素，当 α 与 M 符合一

定条件时，乘同余法产生的伪随机数序列最大可能周期为 2^{k-2}。一般取初值 $x_0=2\alpha+1$ 型的数，$\alpha=8q\pm3$，其中 q 为正整数。最好取 $\alpha=5^{2l+1}$，其中 l 是使 $5^{2l+1}<2^k$ 成立的最大正整数。

2. 混合同余法

混合同余法产生伪随机数序列的递推同余式为

$$\begin{cases} x_{n+1} \equiv \alpha x_n + \beta \quad (\text{mod } M) \\ r_{n+1} = x_{n+1}/M \end{cases} \tag{8-4}$$

混合同余法与乘同余法的差别仅是增加了一个增量 β，其他含义与式(8-3)相同。

混合同余法所产生的伪随机数序列的周期及统计性质与 x_0,α,β,M 的取值有着密切的联系。若取值不当，产生的伪随机数序列的周期可能很短，或者是非随机性的数据序列。一个混合同余伪随机数序列，达到周期为 M 的充分必要条件是：

① β 与 M 互素；

② 对每一个 M 的素因子 p，$\alpha-1$ 为 p 的倍数；

③ 若 M 是 4 的倍数，那么 $\alpha-1$ 为 4 的倍数；

若计算机字长为 k，当 $M=2^k$ 时，应取 $\alpha=4q+1$，$\beta=2\alpha+1$，x_0 为任意的非负整数，其中 q 为正整数。

当 $M=2^{19}=524\ 288$，$\alpha=5^5=3\ 125$ 时，$x_0=23,11,19,37$；$\beta=3,7,11,17$ 分 4 套配合使用，混合同余法产生周期为 524 288 的伪随机数序列。

二、伪随机数序列的统计检验

前已述及，即便是产生伪随机数比较好的同余法，也不能在计算机上产生[0,1]区间上均匀分布的真随机数序列，只能是在参数选取恰当的条件下产生统计性质符合要求、周期长度足够大的伪随机数序列。这就是说，对于选定参数下所产生的一个伪随机数序列，只有通过严格的统计检验后，才能确定它是否可做[0,1]区间上均匀分布的随机数序列使用。一般来说，伪随机数序列的均匀性和独立性(随机性)是需要进行统计检验的主要性质。当然，在解决不同类型的实际问题时，对各种统计性质的要求不同，就要根据要求进行不同的检验。这里仅介绍检验伪随机数序列均匀性和独立性的几种方法。

1. 均匀性检验

均匀性检验包括矩检验、频率检验和累积频率检验。

(1) 矩检验。

矩检验也称参数检验。它是对伪随机数序列各阶矩统计量的一种显著性检验。若伪随机数序列的长度(序列内伪随机数的个数)为 n，那么它的各阶矩为：

$$m_k = \frac{1}{n}\sum_{i=1}^{n} r_i^k$$

式中 m_k——伪随机数序列的 k 阶矩；

r_i^k——[0,1]区间上分布的第 i 个伪随机数的 k 次方。

二阶中心矩(方差)为：

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - m_1)^2 = m_2 - m_1^2$$

在理论上，标准均匀分布的 k 阶矩 μ_k 和总体的方差 σ^2 分别为：

$$\mu_k = 1/(k+1); \quad \sigma^2 = \mu_2 - \mu_1^2 = 1/12$$

假设生成的为伪随机数符合标准均匀分布，那么 m_k, s^2 应与 μ_k, σ^2 一致。

若以 h 个伪随机数为一组，共算出 v 组一阶矩 m_{1j}、二阶矩 m_{2j}、二阶中心矩 s_j^2，它们的平均值为：

$$\bar{m}_1 = \frac{1}{v}\sum_{j=1}^{v} m_{1j}; \quad \bar{m}_2 = \frac{1}{v}\sum_{j=1}^{v} m_{2j}; \quad \bar{s}^2 = \frac{1}{v}\sum_{j=1}^{v} s_j^2$$

根据中心极限定理，当 $v \to \infty$ 时，$\bar{m}_1, \bar{m}_2, \bar{s}^2$ 的分布趋于标准正态分布，其平均值分别趋于 1/2，1/3，1/12，而总体方差分别趋于 $1/(12h)$，$4/(45h)$，$1/(180h)$。因而可建立统计量：

$$u_1 = (\bar{m}_1 - 1/2)\Big/\sqrt{\frac{1}{12hv}},\ u_2 = (\bar{m}_2 - 1/3)\Big/\sqrt{\frac{4}{45hv}},\ u_3 = (\bar{s}^2 - 1/12)\Big/\sqrt{\frac{1}{180hv}}$$

u_1, u_2, u_3 近似服从正态分布，并且可以确定均匀性假设的临界域 R，即

$$R_{\bar{m}_1}:\left(\frac{1}{2} - u_\alpha\sqrt{\frac{1}{12hv}},\ \frac{1}{2} + u_\alpha\sqrt{\frac{1}{12hv}}\right)$$

$$R_{\bar{m}_2}:\left(\frac{1}{3} - u_\alpha\sqrt{\frac{4}{45hv}},\ \frac{1}{3} + u_\alpha\sqrt{\frac{4}{45hv}}\right)$$

$$R_{\bar{s}^2}:\left(\frac{1}{12} - u_\alpha\sqrt{\frac{1}{180hv}},\ \frac{1}{12} + u_\alpha\sqrt{\frac{1}{180hv}}\right)$$

当 $\bar{m}_1, \bar{m}_2, \bar{s}^2$ 大于对应的 R 的上界或小于 R 的下界时，应否认均匀性假设。u_α 可以从正态分布的双侧分位数表中查得。

（2）频率检验。

这种检验又称拟合优度检验。如果把[0,1]区间分成 k（一般取 $k=8,16,32$）个等子区间，那么频率检验即是检验每个子区间的观测频数 n_i 与理论频数 m_i（$m_i = n/k$）之间的差异。为此，可建立自由度为 $k-1$ 的 χ^2 分布检验统计量：

$$\chi^2 = \frac{k}{n}\sum_{i=1}^{k}(n_i - n/k)^2$$

式中　n——被检验的伪随机数个数；

假设 H_0：观测频数与理论频数差异不显著。

给出检验水平 $\alpha = 0.05$，若 $\chi^2 \geqslant \chi^2_{0.05}$，则否定假设 H_0，即 n_i 与 m_i 差异显著（伪随机数序列分布不均匀），反之接受假设 H_0，即伪随机数序列分布均匀。

（3）累积频率检验。

该检验也称柯尔莫果罗夫拟合优度检验。设 $F(r)$ 是伪随机数序列的分布函数，而 $S_n(r)$ 是对该序列进行 n 次独立观测得到的经验分布函数。根据柯尔莫果罗夫-斯米尔诺夫定理，对于任意 $\lambda > 0$，有等式：

$$\lim_{n\to\infty} Q_n(\lambda) = \lim_{n\to\infty} p(D_n < \lambda/\sqrt{n}) = Q(\lambda)$$

式中　$D_n = \sup\limits_{0\leqslant r\leqslant 1} |F(r) - S_n(r)|$，$Q(\lambda) = \sum\limits_{k=-\infty}^{+\infty}(-1)^\lambda e^{-2k^2\lambda^2}$

因此，当试验次数 n 足够大时，就可以认为 $D_n < \lambda/\sqrt{n}$ 的概率 $p(D_n < \lambda/\sqrt{n})$ 趋于 $Q(\lambda)$。

如果 D_n^0 是 n 次试验的 $|F(r) - S_n(r)|$ 中最大者，并且 $\lambda_0 = \sqrt{n}D_n^0$，当

$$p(\sqrt{n}D_n \geqslant \lambda_0) = 1 - Q(\lambda_0) = \alpha$$

很小时，就发生了小概率事件，由此即可检验伪随机数序列的均匀性。当 $\lambda_0 \geqslant 0.05$ 时，则可认为伪随机数序列的不均匀性是显著的，否则认为产生的是均匀分布的伪随机数序列。

2. 独立性检验

它是对伪随机数的自相关性进行的统计检验。检验方法有很多种，在此仅介绍简单的独立性检验和顺序检验。

(1) 简单独立性检验。

简单独立性检验又称无重复列联检验。进行这种检验，首先是把被检验的伪随机数序列等分成 u 与 v 两部分，并且要求任一个伪随机数只能唯一地属于 u 或 v。

设 u,v 两部分的值分别为 $u_t,v_t(t=1,2,\cdots)$，把单位正方形划分为 k 行 h 列的网格，并把点 (u_t,v_t) 落在网格 (i,j) 内的观测频数记为 n_{ij}。若 u,v 相互独立，那么 u_t,v_t 同时出现的概率为：

$$p(u_t,v_t) = p(u_t)p(v_t) \tag{8-5}$$

简单独立性检验就是以式(8-5)为假设 H_0，比较观测频数 n_{ij} 与理论频数 m_{ij} 之间的差异是否显著的一种检验。记

$$n_{i.} = \sum_{j=1}^{h} n_{ij}, \quad n_{.j} = \sum_{i=1}^{k} n_{ij}, \quad n = \sum_{i,j} n_{ij}$$

构造检验统计量

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{n_{ij}^2}{m_{ij} - n}$$

因为

$$p(u_t) \approx n_{i.}/n,\ p(v_t) \approx n_{.j}/n,$$

所以

$$m_{ij} = np(u_t,v_t) \approx n_{i.}n_{.j}/n$$

又因

$$n_{i.} = \sum_{j=1}^{h} n_{ij} = \sum_{j=1}^{h} m_{ij} \qquad (i = 1,2,\cdots,k)$$

$$n_{.j} = \sum_{i=1}^{k} n_{ij} = \sum_{i=1}^{k} m_{ij} \qquad (j = 1,2,\cdots,h)$$

所以 χ^2 的自由度为：

$$hk - (h + k - 1) = (h - 1)(k - 1)$$

假设 $H_0: p(u_t,v_t) = p(u_t)p(v_t)$。

对于检验水平 $\alpha = 0.05$，如果 $\chi^2 < \chi^2_{0.05}$，则可认为 u,v 是相互独立的；若 $\chi^2 \geqslant \chi^2_{0.05}$，则称 u,v 之间相关显著；若 $\chi^2 \geqslant \chi^2_{0.01}$，则称 u,v 之间相关极显著。

(2) 顺序检验。

顺序检验又称有重复列联检验。对于随机性的伪随机数序列，不会出现在某一类型的伪随机数之后总是出现另一类型伪随机数的现象。把单位正方形划分为 k 行 h 列的网格，如果把伪随机数序列中相应的两个伪随机数组成一个点，那么被检验的 n 个伪随机数构成的 n 个点落入所有网格内的频数应近似相等，并且 $n_i = n_j$。若 m_{ij} 是观测频数 n_{ij} 的理论频

数，则有统计量

$$\chi^2 = \delta_2^2 - \delta_1^2$$

式中 $\delta_1^2 = \sum_{i=1}^{k}(n_{i.} - m)^2/m, m = n/h$，自由度为 $h-1$；

$\delta_2^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij} - m)^2/m, m = n/hk$，自由度为 $hk-1$。

为了计算 δ_1^2,δ_2^2，需要把 n 个伪随机数按大小分为 N 种类型，即

$$(j-1)/N \leqslant r_i < j/N \qquad (j = 1,2,\cdots,N)$$

$$(k-1)/N \leqslant r_{i+1} < k/N \qquad (k = 1,2,\cdots,N)$$

分类后便可计算 δ_1^2,δ_2^2 以及 χ^2。对于给定的检验水平 $\alpha=0.05$，若 $\chi^2 \geqslant \chi^2_{0.05}$，则称伪随机数序列显著顺序相关，否则认为伪随机数序列顺序不相关。

§3　随机变量的抽样

在[0,1]区间上均匀分布随机数的基础上，就可以实现对随机变量的随机抽样了。对随机变量的抽样有多种方法，在此介绍经验分布函数抽样法、直接抽样法和变换抽样法。

一、经验分布函数抽样法

1. 随机变量的经验分布函数

按数学中的定义，随机变量 X 的分布函数是 X 的取值不大于某实数 x 的概率 $p(X\leqslant x)$，通常将其记为：

$$F(x) = p(X \leqslant x)$$

随机变量 X 的经验分布函数是由 X 的 n 次观测值 $x_i(i=1,2,\cdots,n)$ 用频率统计法求出的分布函数，并将其记为：

$$F_n(x) = p(X \leqslant x)$$

在油气资源评价及其他地质研究工作中，人们常常需要知道随机变量 X 的取值大于某实数 x 的概率 $p(X\geqslant x)$，显然

$$p(X > x) = 1 - p(X \leqslant x)$$

记

$$F'_n(x) = p(X > x) = 1 - p(X \leqslant x) = 1 - F_n(x) \tag{8-6}$$

式(8-6)是油气资源评价及其他地质研究中常用的随机变量的经验分布函数(图 8-2)。

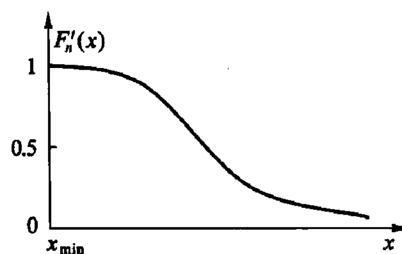


图 8-2　经验分布函数示意图

2. 经验分布函数的构造方法

(1) 频率统计法。

假设已有随机变量 X 的 $n(n\geqslant30)$ 个观测值 $x_i(i=1,2,\cdots,n)$，那么用频率统计法求经验分布函数式(8-6)的大致步骤如下：

① 确定频率统计区间数。

设观测值的个数为 n，区间数为 m(m 为奇数)，并按照 $n/m\geqslant5$ 的原则确定 m 的值。

② 计算区间间隔值。

$m+1$ 个区间间隔值为：

$$x_i = x_{min} + [(x_{max} - x_{min})/m](i-1) \qquad (i = 1,2,\cdots,m+1)$$

式中 x_{max}, x_{min}——分别为随机变量观测值的最大值和最小值。

③ 求经验分布函数。

记 n_i 为观测值落入区间(x_i, x_{i+1})内的频数，f_i 为累加频率，则

$$f_i = \frac{1}{n}\sum_{j=1}^{m} n_j \qquad (i = 1,2,\cdots,m)$$

而经验分布函数为：

$$F'_n(x) = \begin{cases} 1, & x_1 \leqslant x < x_2; \\ f_2, & x_2 \leqslant x < x_3; \\ \vdots & \\ f_m, & x_{m-1} \leqslant x < x_m; \\ 0, & x_m \leqslant x \end{cases}$$

(2) 等概率统计法。

当随机变量的观测值为10～30个，而且又不知道随机变量的分布概率模型时，如果采用频率统计法求随机变量的经验分布函数，就会因统计区间个数少而使得经验分布函数过于粗糙。此时，可视随机变量观测值出现的概率是相等的。如果样本容量为 n，把观测值按大小顺序排列成：

$$x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$$

若 $x_k \leqslant x \leqslant x_{k+1}$，则不大于 x 的观测值的频率为 k/n，那么经验分布函数为：

$$F'_n(x) = 1 - F_n(x)$$

式中

$$F_n(x) = \begin{cases} 0, & x < x_1 \\ \dfrac{k}{n}, & x_k \leqslant x \leqslant x_{k+1} \\ 1, & x_n \leqslant x \end{cases}$$

3. 经验分布函数的抽样

经验分布函数图(图 8-3)的坐标原点为$(x_{min}, 0)$，以随机变量观测值的最小值 x_{min} 与第 i 个伪随机数 r_i 组成数据对，在分布函数图的纵轴上确定一个入口点(x_{min}, r_i)，过该点作横轴的平行线交分布曲线于点(x_i, r_i)，交点的横坐标 x_i 对应于伪随机数 r_i 对随机变量的一次随机抽样值。在经验分布函数抽样中，常称 r_i 为随机抽样的随机入口值或概率入口值，而称 x_i 为随机抽样的出口值。用上述方法可以得到随机变量的一系列随机抽样值 。实际计算时，是用插值(线性或非线性插值)的方法求出伪随机数 r_i 对随机变量的随机抽样值 x_i。

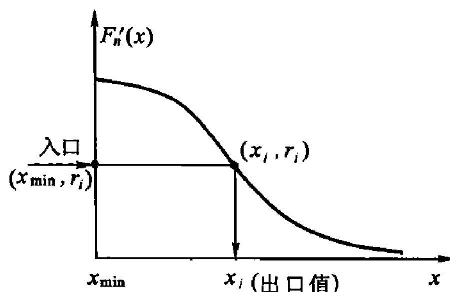


图 8-3 分布函数抽样示意图

二、直接抽样法

直接抽样法是对具有单调递增的连续分布函数，并且又可用区间[0,1]上均匀分布的随

机变量 ζ 的显式表示的随机变量 η 的随机抽样。

在区间[0,1]上均匀分布的随机变量与其他分布的随机变量在理论上有以下重要关系：

假设随机变量 η 具有单调递增的连续分布函数 $F(x)$（或已给出分布密度函数 $f(x)$），则随机变量 $\zeta=F(\eta)$ 是区间[0,1]上均匀分布的随机变量。换句话说，如果 ζ 为区间[0,1]上均匀分布的随机变量，而 $F(x)$ 是某个随机变量的分布函数，并且 $F(x)$ 为单调递增的连续函数，那么 $\eta=F^{-1}(\zeta)$ 是以 $F(x)$ 为分布函数的随机变量。

由上述关系可知，只要随机变量具有连续单调递增的分布函数，$F^{-1}(x)$ 又能够用显式表示出来，就可以用均匀分布的随机抽样序列产生其他分布的随机抽样序列。但一般说来，η 不能用 ζ 的显式写出来，因此，随机变量的直接抽样法仅适合对某些分布概率模型的随机变量抽样。下面是用这种抽样方法抽样的例子。

【例 1】试由区间[0,1]上均匀分布的随机变量 ζ 产生 $[a,b]$ 区间上均匀分布的随机变量 η。

解：$[a,b]$ 区间上均匀分布的随机变量 η 的密度函数为：

$$f(x)=\begin{cases}1/(b-a) & (a\leqslant x\leqslant b)\\ 0 & (\text{其他})\end{cases}$$

根据随机变量间的理论关系，有：

$$\zeta=\int_a^\eta[1/(b-a)]\mathrm{d}x=(\eta-a)(b-a)$$

由此得：

$$\eta=(b-a)\zeta+a$$

【例 2】试由[0,1]区间上均匀分布的随机变量 ζ 产生服从三角分布的随机变量 η。

解：当随机变量的密度函数为 $f(x)=2x(0\leqslant x\leqslant 1)$ 时，称之为三角分布。根据随机变量间的理论关系，有：

$$\zeta=\int_0^\eta 2x\mathrm{d}x=\eta^2$$

由此得：

$$\eta=\zeta^{1/2}$$

结果表明，[0,1]区间上均匀分布的随机变量的平方根服从三角分布。

【例 3】设 ζ 是[0,1]上均匀分布的随机变量，η 是服从指数分布的随机变量，其分布函数为：

$$F(x)=1-\mathrm{e}^{-\lambda x}\qquad(x>0,\lambda\ \text{为常数})$$

试用 ζ 把 η 表示出来。

解：由随机变量间的理论关系可得：

$$\zeta=1-\mathrm{e}^{-\lambda\eta}$$

所以

$$\eta=-\ln(1-\zeta)/\lambda$$

因 ζ 在[0,1]上是均匀分布，而 $1-\zeta$ 在[0,1]上也是均匀分布，故有：

$$\eta=-(\ln\zeta)/\lambda$$

【例 4】试证标准正态分布的随机变量 η 的抽样序列不能用[0,1]上均匀分布的随机变量 ζ 的抽样序列表示出来。

解：根据随机变量的理论关系，由标准正态分布的密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

虽然可得：

$$\zeta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{x^2}{2}} dx$$

但是，η 却不能用 ζ 的显函数形式表示出来，因此，无法由[0,1]上均匀分布的随机变量 ζ 的抽样序列产生标准正态分布随机变量的抽样序列。

三、变换抽样法

变换抽样法是对具有单调递增的连续分布函数，但却不能用[0,1]区间上均匀分布的随机变量 ζ 的显式表示的随机变量 η 的随机抽样。

正态分布是常用的一类分布，它的抽样方法是统计试验法的重要内容。产生服从 $N(0,1)$ 的抽样值 ζ，又是获得服从 $N(a,\sigma^2)$ 的抽样值 η 的基础。

假设 x_i, x_{i+1} 是相互独立的在[0,1]上均匀分布的两个随机数，做如下变换：

$$\begin{cases} \zeta_i = (-2\ln x_i)^{1/2} \cos 2\pi x_{i+1} \\ \zeta_{i+1} = (-2\ln x_i)^{1/2} \sin 2\pi x_{i+1} \end{cases}$$

那么，ζ, ζ_{i+1} 就是两个相互独立的服从 $N(0,1)$ 分布的随机数。

设 $x_1, x_2, \cdots, x_n$ 是 n 个相互独立的[0,1]上均匀分布的随机数，那么 x_i 的期望和方差为：

$$E(x_i) = 1/2, \qquad D(x_i) = 1/12$$

根据中心极限定理，当 n 充分大时，

$$\zeta_n = \left(\sum_{i=1}^{n} x_i - n/2\right) \Big/ \sqrt{n/12}$$

的分布渐近于 $N(0,1)$ 分布，故可把 ζ_n 近似看做服从标准正态分布的随机数。通常取 n 等于 8 或者 12，当 $n=12$ 时变换最方便。

如果随机变量 η 的数学期望为 $E(\eta)$，方差为 $D(\eta)>0$，那么随机变量

$$\zeta = [\eta - E(\eta)]/\sqrt{D(\eta)}$$

服从 $N(0,1)$ 分布，因此有：

$$\eta = \sqrt{D(\eta)} \cdot \zeta + E(\eta)$$

上式表明，在已知随机变量 η 的数学期望和方差的条件下，可由服从 $N(0,1)$ 分布的随机变量直接产生随机变量 η 的随机数。

§4 蒙特卡罗法估算油气资源量

一、油气资源量概率模型

局部含油气地质单元是估算油气资源量的基本地质体。在估算资源量时，它的含义可以不同，既可以是生油凹陷中的一个生油层系，又可以是次一级构造单元中的生油层系，还可以是局部构造等。基于不同的找油理论，资源量的估算方法也不一样，但是任何一个局部含油气地质单元的油气资源量却都是与油气资源量相关的地质参数（常数和变量）的连乘积，即油气资源量估算概率模型的通式为：

$$Q_j = K_j \prod_{i=1}^{n} X_{ji} \tag{8-7}$$

式中　Q_j——含油气区内第 j 个局部地质单元的油气资源量；

K_j——第 j 个局部地质单元内与油气资源量有关的地质常数与经验系数的积；

X_{ji}——第 j 个局部地质单元内与油气资源量有关的第 i 个随机变量。

估算局部含油气地质单元的油气资源量，实际上就是求式(8-7)的数学期望，即

$$E(Q_j) = K_j E(\prod_{i=1}^{n} X_{ji}) \tag{8-8}$$

由式(8-8)可知，当 K_j 中包含局部含油气地质单元的面积时，估算局部含油气地质单元油气资源量的关键是求单位面积的资源量。

二、油气资源量的估算

1. 最大、最小可能资源量及累积频率区间间隔值

为了求资源量的分布曲线(分布函数)，在资源量估算之前，先求出第 j 个局部地质单元资源量的最大、最小可能值和累积频率区间间隔值。

(1) 最大可能值。

$$q_{j\max} = K_j \prod_{i=1}^{n} X_{ji\max} \qquad (j = 1,2,\cdots,m)$$

(2) 最小可能值。

$$q_{j\min} = K_j \prod_{i=1}^{n} X_{ji\min} \qquad (j = 1,2,\cdots,m)$$

(3) 频率区间间隔值。若累积频率区间数为 v，那么区间间隔值为：

$$q_{jh} = q_{j\min} + [(q_{j\max} - q_{j\min})/v](h-1) \qquad (h = 1,2,\cdots,v+1)$$

2. 油气资源量的估算

(1) 随机抽样的插值计算。

如果采用经验分布函数抽样法对随机变量 X_{ji} $(i=1,2,\cdots,n)$ 进行随机抽样，则需要在计算机内存储 X_{ji} 分布曲线上的 u 个控制点 $(x_{jip},F'(x_{jip}))$ $(p=1,2,\cdots,u)$。当抽样点 (x_{jig},r_k) 落入控制点 $(x_{jip},F'(x_{jip}))$ 与 $(x_{jip+1},F'(x_{jip+1}))$ 之间时(图 8-4)，随机变量的抽样值用插值法(线性或非线性)求出。若按线性插值计算，那么第 g 次的抽样值 x_{jig} $(g=1,2,\cdots,N)$ 为：

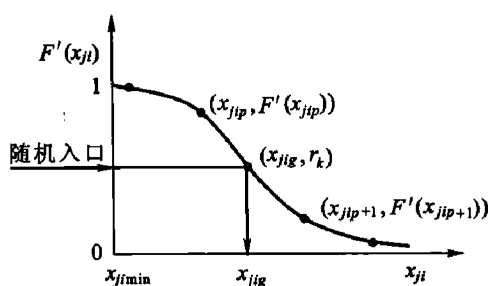


图 8-4　线性插值示意图

$$x_{jig} = x_{jip} + (x_{jip+1} - x_{jip})[r_k - F'(x_{jip})]/[F'(x_{jip+1}) - F'(x_{jip})]$$

$$(j = 1,2,\cdots,m; i = 1,2,\cdots,n; p = 1,2,\cdots,u; g = 1,2,\cdots,N)$$

式中　r_k——伪随机数序列中的第 k 个伪随机数。

(2) 油气资源量估算。

把随机变量 X_{ji} 的抽样值 x_{jig} $(i=1,2,\cdots,n; g=1,2,\cdots,N)$ 代入式(8-7)，得到第 j 个局

部地质单元油气资源量的估计值：

$$q_{jg}=K_j\prod_{i=1}^{n}x_{jig}\qquad(g=1,2,\cdots,N)$$

(3) 第 j 个局部地质单元油气资源量分布函数。

根据第 j 个局部地质单元油气资源量区间间隔值 q_{jh}，用频率统计方法求出 q_j 的分布函数 $F(q_j)$（图 8-5）。

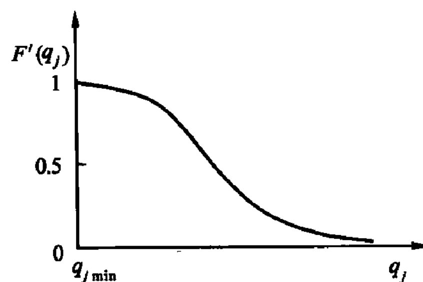


图 8-5　资源量分布函数示意图

3. 油气资源总量

如果含油气区共有 m 个局部含油气地质单元，那么含油气区的油气资源总量为 m 个油气资源量的概率和：

$$q=\sum_{j=1}^{m}q_j$$

概率求和的大致过程如图 8-6 所示。

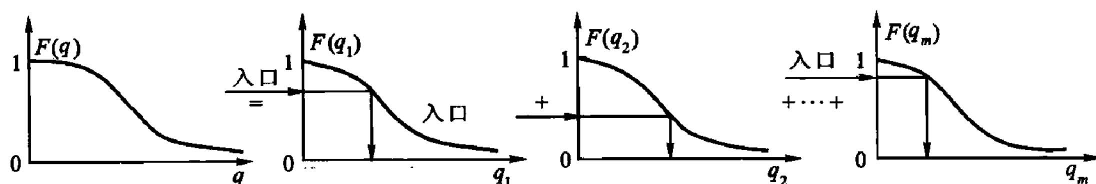


图 8-6　油气资源量的概率和示意图

三、地质风险分析

由于石油勘探的未来成效具有不确定性，因而需要对估算的石油资源量进行风险分析。石油勘探中的风险是多种多样的，如勘探区是否具备形成油气藏的地质条件的风险；在具备形成油气藏地质条件的含油气区内，能否找到一定规模的油气藏的勘探风险；勘探后发现的油气藏是否具备开采价值的经济风险；勘探过程中人与设备是否安全的环境风险；对于勘探地区，特别是大陆架地区是否有国际争议的政治风险等。

油气资源评价为油气勘探提供决策依据，而油气资源估算是资源评价的组成部分。因此，地质人员要对上述风险下估算的油气资源量做好地质风险分析。在实际的工作中，地质风险分析可以在不同的层次进行，如单一地质圈闭的风险分析，一组地质圈闭（国外将地质条件相似的一组地质圈闭称做一个勘探层）的风险分析，一个油气聚集带的风险分析，整个含油气盆地的风险分析等。

地质风险分析大多从圈闭做起，其风险的计算公式为：

$$K=1-\prod_{i=1}^{n}(1-k_i)\tag{8-9}$$

式中　K——圈闭风险值；

　　k_i——第 i 个因素的风险值。

例如，若用容积法估算圈闭的石油储量，则有：

$$Q=S\cdot H\cdot\varphi\cdot D\cdot W\tag{8-10}$$

式中　Q——石油储量；

　　S——含油面积；

　　H——储集层厚度；

φ——储集层孔隙度；

D——石油充满系数；

W——采收率。

对储量进行风险分析，就要由熟悉含油气区地质资料的人员对上式等号右边的五项参数逐个分析论证。通常来说，S 的风险常常取决于地质调查或地震勘探资料的可靠性；H 的风险受岩性岩相变化的影响；φ 的风险取决于储集层孔隙是否有次生改造或后期充填的影响；D 的风险受生油岩的成熟度和油气运移通道的制约；W 的风险则与原油性质及驱动类型有关。

经过风险分析论证，对每个参数给出风险值 k ($0\leqslant k\leqslant 1$)。目前尚无完善的方法确定风险值，一种方法是由分析论证人凭经验指定；另一种方法是借用地质条件类似的临区的风险值。例如，某个地质圈闭经风险分析论证后给出表 8-1 所示风险值，根据表内数据，按式(8-9)计算该圈闭的风险值 $K=0.65$，而保险值为 0.35，即经过风险分析后该地质圈闭的石油储量仅是风险分析前的 35%(图 8-7)。

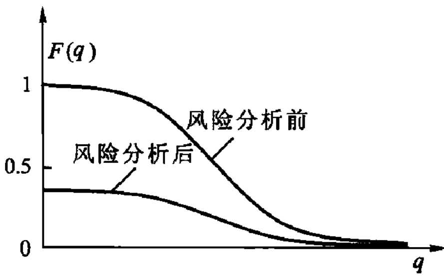


图 8-7　风险分析前后石油储量关系

表 8-1　单一地质圈闭风险数据

地质参数	风险值 k	保险值$(1-k)$
含油面积	0.0	1.0
储集层厚度	0.5	0.5
储集层孔隙度	0.0	1.0
石油充满系数	0.3	0.7
采收率	0.0	1.0

如果有一组(10 个)在地质条件上相类似的可能含油气圈闭，按式(8-10)估算它们的石油储量，经过分析论证，给出地质圈闭的各项参数风险值(表 8-2)，那么在该组地质圈闭中获得一个油藏的可能性有多大？

表 8-2　一组可能含油圈闭的风险数据

地质参数 \ $1-k$ \ 圈闭序号	1	2	3	4	5	6	7	8	9	10
S	1.0	1.0	0.5	1.0	0.5	1.0	1.0	1.0	1.0	1.0
H	0.5	1.0	1.0	1.0	1.0	0.5	1.0	0.5	1.0	1.0
φ	1.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5
D	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
W	1.0	1.0	1.0	0.5	1.0	1.0	0.5	1.0	0.5	1.0

对于这个问题有两种计算方法：

① 考虑所有地质参数的影响，在该组可能含油的地质圈闭中发现一个油藏的风险值

$$K=\prod_{j=1}^{10}\left[1-\prod_{i=1}^{5}(1-k_{ji})\right]=(1-0.25)^{10}=0.0563$$

而在该组可能含油的地质圈闭中发现一个油藏的保险值

$$P = 1 - K = 1 - 0.0563 = 0.9437$$

即发现一个油藏的可能性为94.4%。

② 仅考虑最不利的地质参数的影响，在该组可能含油的地质圈闭中发现一个油藏的风险值

$$K = \prod_{j=1}^{10} k_i = (1-0.5)^{10} = 0.00098$$

而在该组可能含油的地质圈闭中发现一个油藏的保险值

$$P = 1 - K = 0.99902$$

即发现一个油藏的可能性为99.9%。

进行地质风险分析时，如果地质参数间具有多层次结构，则应计算复合地质风险值。例如，探区的地质风险取决于生油和储油条件，而生油条件与生油层厚度及生油相带有关；储油条件与储层厚度及储层相带有关。经过分析论证，给出复合地质风险值(表8-3)。

表8-3 复合地质风险值

基础地质因素	风险值 k_{ij}	保险值 $1-k_{ji}$	复合地质因素	风险值 k_j	保险值 $1-k_j$
生油层厚度	0.4	0.6	生油条件	0.4	0.6
生油层相带	0.0	1.0			
储集层厚度	0.1	0.9	储油条件	0.37	0.63
储集层相带	0.3	0.9			

复合地质风险值可按下式计算：

$$K = 1 - \prod_{j=1}^{m} \{1 - [1 - \prod_{i=1}^{n}(1 - k_{ji})]\} = 1 - \prod_{j=1}^{m}\prod_{i=1}^{n}(1 - k_{ji}) \quad (8\text{-}11)$$

式中 K——复合风险值；

k_{ji}——第 j 项复合地质因素的第 i 个基础地质因素的风险值。

根据表8-3中的数据，按式(8-11)计算，复合地质因素风险值

$$K = 1 - (0.6 \times 0.1) \times (0.9 \times 0.7) = 0.9622$$

在本例中，地质数据结构为两层，当结构层次多于两层时，复合地质风险值的计算将复杂一些。

四、风险分析后资源总量的估算

若含油气区内第 j 个局部含油气地质单元油气资源量的分布函数为 $F(q_j)$，地质风险值为 k_j，保险值为 $1-k_j$，那么在重新估算含油气区地质风险分析后的油气资源总量时，随机抽样入口值的分布区间应由[0,1]变为$[0,1-k_j]$（图8-7）。因此，应将[0,1]上分布的随机数 r_s 改造为$[0,1-k_j]$上分布的随机数 r_s^*，且 $r_s^* = (1-k_j)r_s$。

在以上基础上，仍然用插值的方法求出风险分析后 q_j 的抽样值 $q_{j\alpha}(j=1,2,\cdots,m;\alpha=1,2,\cdots,N)$，再由 $q_{j\alpha}$ 求得油气资源总量估计值 $q_\alpha(\alpha=1,2,\cdots,N)$，最后利用频率统计法求出地质风险分析后资源总量的分布函数 $F(q)$。

§5 应用实例

某沉积盆地中的一个地质凹陷有三套生油层系。用氯仿沥青法估算该凹陷的远景石油

资源量，各生油层系的石油资源量估算公式为：

$$Q_j = S_j \cdot H_j \cdot D \cdot A_j \cdot k_1 \cdot k_2 \qquad (8\text{-}12)$$

式中 Q_j——第 j 套生油层系的石油资源量；

S_j——第 j 套生油岩的分布面积，km^2；

H_j——第 j 套生油岩的厚度，m；

D——生油岩密度，$\times 10^8$ t/km^3；

A_j——第 j 套生油岩氯仿沥青含量，%；

k_1——排烃系数；

k_2——聚集系数。

凹陷石油资源量为三套生油层系资源量的和。三套生油层系的地质参数见表 8-4。

表 8-4 三套生油层系的地质参数

地质参数 \ 生油层系		第一套生油层系	第二套生油层系	第三套生油层系
生油岩分布面积/km^2		1 400	7 000	3 000
生油岩密度/($\times 10^8$ t·km^3)		23	23	23
排烃系数		0.44	0.48	0.43
聚集系数		0.111	0.111	0.111
生油岩厚度/m	数据个数	140	70	30
	取值范围	0.1～1.0	0.1～1.0	0.1～1.0
氯仿沥青含量/%	数据个数	37	37	21
	取值范围	0.06～1.97	0.05～1.54	0.02～1.74

对于某生油层系，其分布面积、密度为常数；排烃系数、聚集系数是经验系数；生油岩厚度、生油岩氯仿沥青含量为有一定取值范围的随机变量，其分布函数如图 8-8 所示。

不同概率下三套生油层系的石油资源量及全凹陷石油资源总量见表 8-5。

表 8-5 各生油层及全凹陷石油资源量汇总表

概率/% \ 生油层位及资源量	第一套生油岩石油资源量/($\times 10^8$ t)	第二套生油岩石油资源量/($\times 10^8$ t)	第三套生油岩石油资源量/($\times 10^8$ t)	全凹陷石油总资源量/($\times 10^8$ t)
100	2.244 5	3.815 9	1.952 7	11.546 8
90	4.562 1	5.878 8	4.998 6	16.668 7
80	5.083 6	6.394 5	5.675 4	17.949 2
70	5.431 2	6.795 6	6.081 5	18.863 8
60	5.720 9	7.196 7	6.487 6	19.595 5
50	6.010 6	7.483 2	6.758 4	20.144 3
40	6.300 3	7.769 7	7.096 8	20.693 1
30	6.590 0	8.056 2	7.367 5	21.241 8
20	6.879 7	8.342 7	7.570 6	21.973 5
10	7.169 4	8.686 5	7.841 3	22.705 3
0	8.038 5	9.488 7	8.653 6	25.815 0

第一套生油岩氯仿沥青分布函数

第一套生油岩氯仿沥青分布密度

第一套生油岩厚度分布函数

第一套生油岩厚度分布密度

第一套生油岩石油资源量分布函数

第二套生油岩氯仿沥青分布函数

第二套生油岩氯仿沥青分布密度

第二套生油岩厚度分布函数

第二套生油岩厚度分布密度

第二套生油岩石油资源量分布函数

第三套生油岩氯仿沥青分布函数

第三套生油岩氯仿沥青分布密度

第三套生油岩石油资源量分布函数

第三套生油岩厚度分布函数

第三套生油岩厚度分布密度

凹陷石油资源量分布函数

图 8-8 生油岩生油参数及石油资源量分布函数

思考与练习

1. 蒙特卡罗法的现代含义是什么？

2. 对随机变量进行随机抽样时，为何要产生[0,1]上的随机数？

3. 如何构造地质随机变量的经验分布函数？它与概率论中随机变量的分布函数有何不同？

4. 估算油气资源量时，如何实现对随机变量的随机抽样？常用的抽样方法有几种？其各自的适用条件是什么？

5. 蒙特卡罗法估算油气资源量比传统估算方法有何优点？

6. 应用蒙特卡罗法求数值解不确定问题概率解的基本过程是什么？

7. 根据地震资料和少量钻井资料，获得某凹陷下第三系某储集层的厚度数据(表8-6)，试据表中数据求储集层厚度的经验分布函数。

表8-6 储集层厚度数据表

序　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
厚度/m	2.0	4.6	3.9	4.7	2.9	5.1	1.5	5.2	4.2	5.9	4.8	6.3	3.9	6.3	4.7	6.5
序　号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
厚度/m	3.5	4.3	7.1	4.4	5.3	3.0	4.0	7.1	5.4	3.2	6.8	3.7	5.0	4.9	5.5	7.6
序　号	33	34	35	36	37	38	39	40	41	42	43	44	45	46		
厚度/m	3.8	5.6	5.7	2.6	8.0	8.3	5.8	6.2	5.4	6.1	4.1	5.2	5.3	2.9		

第九章 地质数据序列分析

在地质研究中，有很多地质变量的观测值会构成一个有序的数据序列，例如沿地层剖面或钻井岩心剖面上地层的厚度、矿物的成分、电性曲线和模拟地震记录离散抽样、数字地震记录等，这种有序的地质数据描述了地质特征的变化。因此，地质数据序列是描述地质特征的地质变量的观测值按观测顺序排列的数据序列，记为：

$$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$$

地质数据序列分析是研究多个地质数据序列间的相互关系、自身性质等的统计分析方法。在此仅介绍简单的相关分析和滑动平均。

§1 相关分析

相关分析本质上是一种线性滤波，是数字地震资料的一种基本处理方法。它不仅是压制随机干扰、提高讯噪比的一种重要的滤波方法，同时还在识别和消除多次波、计算速度谱、确定同相轴的基本参数等方面都有着广泛的应用。地震勘探中一种独特的方法——连续振动法也是根据相关分析原理提出的。在此，我们把地质数据序列视为一个波形的离散抽样，并从波形的相似出发，引出表示波形相似程度的相关函数，分析相关函数的性质，简单介绍相关分析在地质数据序列研究中的应用。

一、相关函数

对于波形之间的相似性问题，有时从直观上就可以看出“这两个波形比较相似”，“那两个波形不大相似”。例如，波形 $x_1(t)$ 与 $y_1(t)$ 不相似，而 $x_2(t)$ 与 $y_2(t)$ 很相似（图 9-1）。但有时就不一定能从直观上做出这样的定性判断，即使能做出定性判断，那么它们又相似到什么程度呢？下面对此问题进行讨论。

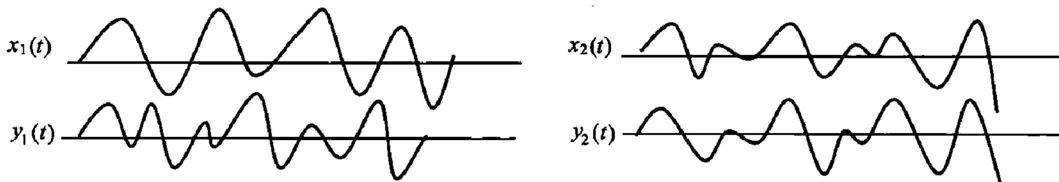


图 9-1 波形的相似性示意图

设 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 是 $x(t)$ 和 $y(t)$ 等间隔的离散抽样值，则它们的均方差

$$\sigma=\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2$$

反映了两个波形的相似程度。σ 越大，$x(t)$ 与 $y(t)$ 的相似程度就越低；反之，$x(t)$ 与 $y(t)$ 的相似程度就越高。把上式展开，得：

$$\sigma=\frac{1}{n}\sum_{i=1}^{n}x_i^2+\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{2}{n}\sum_{i=1}^{n}x_iy_i$$

由上式可知，两个波形的相似程度取决于

$$\varphi = \frac{2}{n}\sum_{i=1}^{n} x_i y_i \tag{9-1}$$

的大小(注意 φ 是可正可负的)。φ 小则 σ 大,两个波形整体的相似程度低;反之,两个波形整体的相似程度高。

在实际工作中,讨论两个波形的相似性时,不仅要从整体上分析它们的相似程度,而且还要分析不同波形上的相似段,即两个波形中的一个相对于另一个移动时,移动到什么位置它们最相似。例如,粗看 $x_3(t)$ 与 $y_3(t)$ 是很不相似的,但仔细观察可以发现 $x_3(t)$ 的 AB 段与 $y_3(t)$ 的 CD 段很相似(图 9-2)。

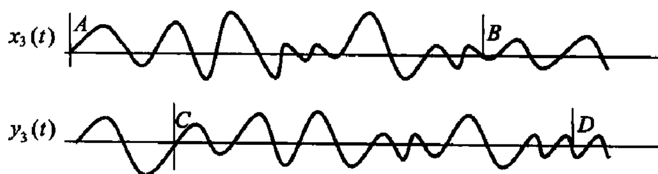


图 9-2　波的相似段

为了确定不同波形上最相似的段,我们计算两个波形相对移动不同位置时的 φ 值,即

$$\varphi(\tau) = \frac{1}{n}\sum_{i=1}^{n} x_i y_{i+\tau} \qquad (\tau = \tau_1, \tau_2, \cdots) \tag{9-2}$$

式中　τ——数据序列的滞后项数。

令 $\tau=\tau_1,\tau_2,\cdots$,可计算出 $\varphi(\tau_1),\varphi(\tau_2),\cdots$,当两个波形相似的段($AB$ 和 CD)恰好重合时,$\varphi(\tau)$最大。由此可知,式(9-2)是定量衡量两个波形相似程度的指标,我们可以利用它确定两个波形中的相似段。另外,$\varphi(\tau)$不仅与两个波形本身的特点有关,而且还与两个波形之间的相对移动量有关。$\varphi(\tau)$称为两个波形的相关函数。

二、互相关和自相关函数

1. 互相关函数

若 $x(t)$,$y(t)$是两个不同的波形,则称式(9-2)为 $x(t)$和 $y(t)$的互相关函数,常记为:

$$\varphi_{xy}(\tau) = \frac{1}{n}\sum_{i=1}^{n} y_i x_{i+\tau} \tag{9-3}$$

$$\varphi_{xy}(\tau) = \frac{1}{T}\int_0^T y(t)x(t+\tau)\mathrm{d}t \tag{9-4}$$

一般来说,互相关函数不是偶函数,$\varphi_{xy}(0)$不一定是最大值。最大值一般在两个波形的最大相似段上。当 τ 趋于无穷大时,$\varphi_{xy}(\tau)$趋于 0。

2. 自相关函数

当 $x(t)=y(t)$时,则称式(9-2)为 $x(t)$的自相关函数,常记为:

$$\varphi_{xx}(\tau) = \frac{1}{n}\sum_{i=1}^{n} x_i x_{i+\tau} \tag{9-5}$$

$$\varphi_{xx}(\tau) = \frac{1}{T}\int_0^T x(t)x(t+\tau)\mathrm{d}t \tag{9-6}$$

自相关函数的图形以 $\varphi_{xx}(\tau)$为对称轴(图 9-3),即 $\varphi_{xx}(\tau)=\varphi_{xx}(-\tau)$,也就是说自相关函数是 τ 的偶函数。当 $\tau=0$ 时,$\varphi_{xx}(0)$为正的最大值;当 τ 趋于无穷大时,$\varphi_{xx}(\tau)$趋于 0。

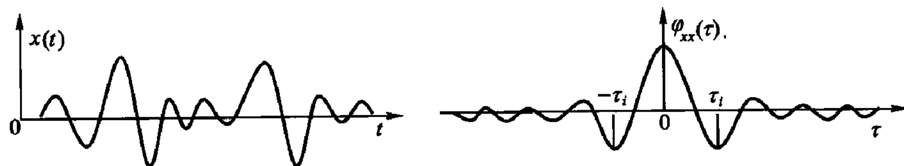


图 9-3 波形及其自相关函数曲线

§2 滑动平均

地质数据序列中的每个观测值均包含趋势(背景)变化、周期性变化和随机干扰三部分，其中对油气勘探有用的信息是趋势和周期性成分。因此，在通过地质数据研究地质问题时，应尽量消除或抑制数据序列中的随机干扰。在此介绍一种消除或抑制随机干扰的简单方法——滑动平均法。下面是常用的几个滑动平均方程：

1. 3 项滑动平均方程

3 项滑动平均是以数据序列中 x_i 为中心，左右各取 1 个相邻数据的加权(权相等)平均，相应的方程为：

$$y_i = (x_{i-1} + x_i + x_{i+1})/3 \qquad (i = 2,3,\cdots,n-1) \tag{9-7}$$

滑动平均过程如下所示：

$$\overbrace{x_1 \quad x_2 \quad x_3}^{y_2} \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad \cdots \quad x_n$$

$$x_1 \quad x_2 \quad \overbrace{x_3 \quad x_4 \quad x_5}^{y_4} \quad x_6 \quad x_7 \quad \cdots \quad x_n$$

$$x_1 \quad \underbrace{x_2 \quad x_3 \quad x_4}_{y_3} \quad x_5 \quad x_6 \quad x_7 \quad \cdots \quad x_n$$

$$x_1 \quad x_2 \quad x_3 \quad \underbrace{x_4 \quad x_5 \quad x_6}_{y_5} \quad x_7 \quad \cdots \quad x_n$$

2. 5 项滑动平均方程

5 项滑动平均是以数据序列中 x_i 为中心，左右各取 2 个相邻数据的加权(权不等，离中心点越远，权越小)平均，相应的方程为：

$$y_i = [17x_i + 12(x_{i+1} + x_{i-1}) - 3(x_{i+2} + x_{i-2})]/35 \qquad (i = 3,4,\cdots,n-2) \tag{9-8}$$

滑动平均过程如下所示：

$$\overbrace{x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5}^{y_3} \quad x_6 \quad \cdots \quad x_{n-1} \quad x_n$$

$$x_1 \quad \underbrace{x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6}_{y_4} \quad \cdots \quad x_{n-1} \quad x_n$$

3. 7 项滑动平均方程

7 项滑动平均是以数据序列中 x_i 为中心，左右各取 3 个相邻数据的加权(权不等，离中心点越远，权越小)平均，相应的方程为：

$$y_i = [7x_i + 6(x_{i+1} + x_{i-1}) + 3(x_{i+2} + x_{i-2}) - 2(x_{i+3} + x_{i-3})]/21$$

$$(i = 4,5,\cdots,n-3) \tag{9-9}$$

滑动平均过程与上类似。

4. 9 项滑动平均方程

9 项滑动平均是以数据序列中 x_i 为中心，左右各取 4 个相邻数据的加权(权不等，离中心点越远，权越小)平均，相应的方程为：

$$y_i = 0.31[7x_i + 0.16(x_{i+1} + x_{i-1}) + 0.08(x_{i+2} + x_{i-2}) - 0.04(x_{i+3} + x_{i-3}) + 0.02(x_{i+4} + x_{i-4})] \quad (i = 5,6,\cdots,n-4) \tag{9-10}$$

滑动平均过程与上类似。

5. 11 项滑动平均方程

11 项滑动平均是以数据序列中 x_i 为中心，左右各取 5 个相邻数据的加权（权不等，离中心点越远，权越小）平均，相应的方程为：

$$y_i = [89x_i + 84(x_{i+1} + x_{i-1}) + 69(x_{i+2} + x_{i-2}) + 44(x_{i+3} + x_{i-3}) + 9(x_{i+4} + x_{i-4}) - 36(x_{i+5} + x_{i-5})]/429 \quad (i = 6,7,\cdots,n-5) \tag{9-11}$$

滑动平均过程与上类似。

6. 15 项滑动平均方程

15 项滑动平均是以数据序列中 x_i 为中心，左右各取 7 个相邻数据的加权（权不等，离中心点越远，权越小）平均，相应的方程为：

$$y_i = [74x_i + 67(x_{i+1} + x_{i-1}) + 46(x_{i+2} + x_{i-2}) + 21(x_{i+3} + x_{i-3}) + 3(x_{i+4} + x_{i-4}) - 5(x_{i+5} + x_{i-5}) - 6(x_{i+6} + x_{i-6}) - 3(x_{i+7} + x_{i-7})]/320 \quad (i = 8,9,\cdots,n-7) \tag{9-12}$$

滑动平均过程与上类似。

7. 21 项滑动平均方程

21 项滑动平均是以数据序列中 x_i 为中心，左右各取 10 个相邻数据的加权（权不等，离中心点越远，权越小）平均，相应的方程为：

$$y_i = [60x_i + 57(x_{i+1} + x_{i-1}) + 47(x_{i+2} + x_{i-2}) + 33(x_{i+3} + x_{i-3}) + 18(x_{i+4} + x_{i-4}) + 6(x_{i+5} + x_{i-5}) - 2(x_{i+6} + x_{i-6}) - 5(x_{i+7} + x_{i-7}) - 5(x_{i+8} + x_{i-8}) - 3(x_{i+9} + x_{i-9}) - (x_{i+10} + x_{i-10})]/350 \quad (i = 11,12,\cdots,n-10) \tag{9-13}$$

滑动平均过程与上类似。

上述各式中，x_i 是原始观测值；y_i 是滑动平均后的新数据；n 为数据序列的长度。

§3 应用实例

【例 1】数据序列的周期性分析。

对美国绿河（Green River）油页岩季候泥厚度数据（表9-1）进行标准差标准化后做自相关分析，并分别以 $\varphi_{xx}(\tau)$ 和 τ 为纵、横坐标绘制自相关图（图 9-4）。由图可以看出，季候泥厚度的变化具有 20 年的周期，这与太阳黑子 22 年的活动周期接近，由此表明，季候泥厚度的变化与太阳黑子的活动有关。

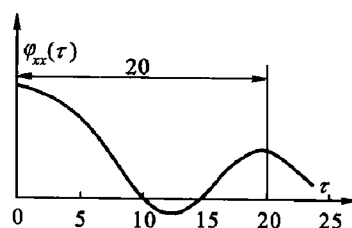


图 9-4 油页岩季候泥分层

表 9-1 绿河油页岩季候泥厚度数据

序　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
厚度/m	6.0	7.2	7.1	7.1	7.2	7.4	8.0	8.6	10.0	11.4	12.0	11.0	9.6	8.7	7.6	7.2
序　号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
厚度/m	7.2	7.8	8.1	7.8	7.1	7.2	7.1	7.0	7.0	7.7	8.6	9.0	12.0	13.7	14.0	13.6
序　号	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
厚度/m	12.0	12.9	12.8	11.1	9.0	7.5	7.5	8.4	7.9	7.0	6.7	6.8	7.3	7.3	7.2	8.1
序　号	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
厚度/m	9.8	11.0	10.8	9.5	8.1	7.2	7.1	6.8	7.0	7.1	5.6	3.8	3.4	4.2	4.8	4.5
序　号	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
厚度/m	5.6	3.0	2.8	4.1	6.8	8.1	7.8	6.4	3.7	4.0	4.2	4.5	5.9	7.5	7.3	6.7
序　号	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	
厚度/m	6.0	5.8	8.7	6.5	13.2	12.4	9.7	9.2	9.3	8.5	6.0	5.7	6.1	6.3	6.3	

【例 2】确定断层断点位置与落差。

有相邻的两口井 No1 和 No2，其中 No1 井通过一条断层(图 9-5a)。对两井的部分对比电测曲线以 1 m 的深度间隔采样，形成数据序列(表 9-2 和 9-3)。

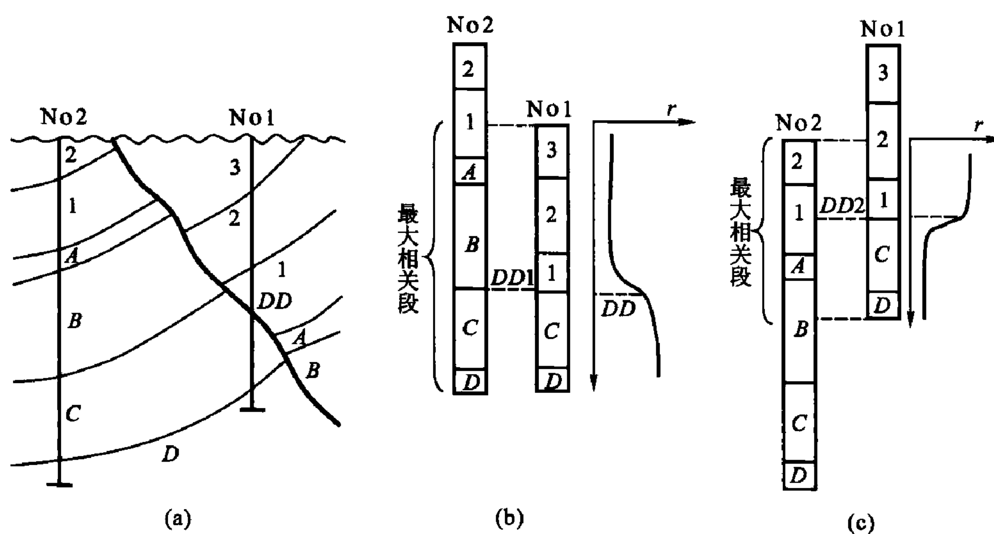


图 9-5 钻井剖面及剖面对比过程示意图

表 9-2 No1 井电测曲线采样值

序　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
μ/mV	2.1	2.3	2.4	2.2	2.2	2.2	2.4	2.2	4.2	9.7	1.9	2.9	2.0	2.2	2.8	1.9
序　号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
μ/mV	2.1	2.0	1.8	1.7	1.8	1.6	1.6	1.5	1.3	1.4	1.6	0.9	1.4	1.9	1.6	1.6

续表

序 号	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
μ/mV	1.5	1.6	1.4	1.5	1.6	1.7	1.6	1.5	1.5	1.5	1.4	1.5	1.7	1.8	2.2	2.8
序 号	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
μ/mV	1.6	2.0	1.9	2.0	1.8	2.4	2.0	2.1	1.8	1.6	1.6	1.7	1.3	1.5	1.4	1.5
序 号	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
μ/mV	2.0	2.1	2.6	3.0	2.1	2.4	1.9	2.2	1.2	2.0	4.0	6.8	5.2	8.6	7.4	6.4
序 号	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
μ/mV	5.3	10.1	4.8	5.4	10.2	6.4	6.5	5.2	4.0	3.5	3.1	4.0	2.0	1.9	2.5	1.8
序 号	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
μ/mV	1.7	3.2	2.0	3.0	3.9	3.6	2.4	2.2	3.1	2.9	2.4	2.0	1.7	2.1	3.1	4.0
序 号	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
μ/mV	3.2	2.2	1.7	1.2	1.3	1.7	5.0	1.2	1.4	3.4	2.4	1.4	1.5	2.3	1.5	2.1
序 号	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
μ/mV	2.1	1.6	2.0	3.5	1.0	2.1	2.3	1.6	2.5	1.3	1.2	2.4	3.4	1.7	5.3	3.8
序 号	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
μ/mV	3.0	1.8	1.2	1.8	1.1	2.6	1.2	1.8	2.7	1.9	1.3	2.4	0.8	0.9	2.1	1.8
序 号	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176
μ/mV	1.7	1.4	1.4	1.8	1.2	1.4	1.6	1.8	1.7	1.8	2.0	1.8	2.5	1.8	1.2	2.8
序 号	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192
μ/mV	2.9	1.8	1.9	3.0	1.8	3.4	2.2	3.2	1.6	1.4	3.2	3.2	2.2	2.0	2.4	2.6
序 号	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
μ/mV	3.6	5.9	3.5	3.6	2.6	9.0	6.0	3.0	13.0	9.0	6.0	13.0	2.2	3.7	4.0	4.6
序 号	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224
μ/mV	4.1	8.8	4.2	2.8	4.0	3.0	2.2	4.6	3.0	2.4	3.2	4.6	2.6	4.0	5.0	2.6
序 号	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
μ/mV	5.3	3.04	4.0	6.0	12.0	17.5	6.0	7.2	4.2	2.0	2.0	3.7	2.6	4.0	4.6	3.0

表 9-3 No2 井电测曲线采样值

序 号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
μ/mV	3.4	3.5	2.8	3.8	2.5	3.1	3.1	2.8	3.0	8.0	6.4	1.6	4.7	2.1	2.6	2.0
序 号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
μ/mV	2.1	2.1	2.0	2.2	2.1	2.0	2.1	2.0	2.5	2.0	1.4	2.4	2.0	1.6	2.0	2.0
序 号	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
μ/mV	2.0	2.1	1.6	1.6	2.0	1.6	1.7	2.1	2.0	2.1	2.6	3.6	4.2	2.4	2.8	2.6

续表

序 号	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
μ/mV	2.2	1.5	2.0	1.8	2.0	1.8	1.8	2.2	2.1	1.6	2.0	.2	1.4	1.5	4.0	3.7
序 号	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
μ/mV	2.3	3.4	6.8	5.2	10.0	20.0	12.4	18.8	2.2	3.3	3.6	2.6	3.6	2.0	1.8	2.4
序 号	80	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
μ/mV	1.8	2.0	1.8	2.0	1.8	2.2	2.6	3.4	2.7	2.5	2.2	1.9	3.0	2.5	2.1	1.7
序 号	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
μ/mV	1.5	2.8	2.0	1.4	1.8	2.3	2.1	2.0	1.4	2.3	2.6	1.6	3.0	2.0	2.2	2.0
序 号	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
μ/mV	1.7	2.2	2.5	2.3	2.8	3.5	2.4	1.6	1.7	2.8	2.6	1.2	2.3	2.4	3.3	2.6
序 号	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
μ/mV	2.1	1.6	2.0	2.8	2.6	1.5	1.4	3.9	2.8	2.3	1.9	3.0	1.6	1.9	4.4	2.5
序 号	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
μ/mV	4.0	2.6	2.0	4.8	3.6	2.0	2.0	3.0	1.4	1.5	8.4	1.1	1.6	1.6	1.4	2.4
序 号	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176
μ/mV	5.2	3.3	4.6	6.6	4.6	4.2	8.4	3.5	6.4	5.4	3.3	3.8	3.8	3.7	3.3	4.0
序 号	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192
μ/mV	2.4	1.8	2.3	1.8	1.8	3.4	1.9	2.6	4.2	3.0	4.4	5.2	2.0	2.2	3.0	2.0
序 号	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
μ/mV	1.8	2.2	3.0	5.0	3.2	2.6	1.8	1.4	1.6	2.2	2.4	2.0	1.3	4.0	7.6	1.0
序 号	209	210	211	212	123	214	215	216	217	218	219	220	221	222	223	224
μ/mV	1.6	2.0	1.7	1.8	3.0	2.0	1.2	1.9	3.1	2.0	1.5	2.3	1.5	2.0	1.8	1.4
序 号	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
μ/mV	1.3	8.0	6.0	0.4	8.8	4.6	3.4	2.0	1.5	1.3	2.0	1.2	2.0	1.2	2.0	2.8
序 号	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256
μ/mV	2.1	1.0	3.0	1.0	1.0	2.2	2.0	2.1	1.5	1.5	1.8	1.6	1.4	1.7	1.8	1.8
序 号	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272
μ/mV	1.9	2.0	2.0	2.0	2.2	2.4	1.8	1.6	3.0	2.8	2.2	1.8	3.2	1.8	2.4	4.4
序 号	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288
μ/mV	1.6	2.3	2.8	2.0	3.2	3.0	2.4	2.4	2.0	2.6	2.0	1.7	9.0	3.0	1.4	2.6
序 号	289	290	291	292	293	294	295	296	297	298	299	300	301	302	302	304
μ/mV	3.0	6.6	1.2	3.0	6.4	3.7	8.0	2.6	2.9	2.4	8.6	2.0	4.0	8.2	3.0	2.2
序 号	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320
μ/mV	7.0	5.0	1.1	9.0	3.6	1.6	2.6	5.0	4.2	2.0	5.4	4.0	2.2	5.3	2.4	4.4

续表

序　号	321	322	323	324	325	326	327	328	329							
μ/mV	3.6	3.2	2.0	2.6	4.5	1.6	1.6	1.6	2.9							

对具有断层、地层尖灭、不整合面等复杂地质现象的地层剖面进行对比时，相邻剖面间会不止一次地出现相似的层段，由此绘出的互相关系数曲线也就有多个峰值，每个峰值对应一个最大(相对)相关段，它隐含着一定的地质意义。在此情况下，就要仔细分析、研究这些最大相关段在地质上的意义。现以具有断层的地层剖面对比为例，说明寻找两个剖面上的最大相关段、确定断点位置和断层落差的过程。

如果把整个对比过程分为 N 段，那么互相关系数曲线会出现 N 个峰值，在每段内找出峰值对应的最大相关段，由此可以找出 N 个最大相关段，其中包括处于图 9-5b 及 9-5c 位置时的两个最大相关段。现在分析这两个最大相关段的特点：当对比段处在图 9-5b 的位置时，对比段明显地分为两部分，下部地层相同，即 No1 井断层的上升盘地层与 No2 井相对地层是同层；上部地层不同，即 No1 井断层的下降盘地层与 No2 井相对地层不是相同的地层。根据上述特点，计算该段(图 9-5b 段)上的累加相关系数公式为：

$$\varphi(i)=\frac{1}{n_i}\sum_{k=1}^{n_i}y_{ik}x_{ik} \tag{9-14}$$

式中　i——累加相关段号；

n_i——第 i 个累加相关段内数据的个数；

y_{ik}，x_{ik}——第 i 个累加相关段内 y，x 的第 k 个采样值。

计算累加相关系数时，最大相关段中的两个数据序列不再移动，仅是从上部开始，依次增大计算累加相关系数的累加长度。从开始到断点 DD，因为最大相关段上部两个剖面对应的地层不同，所以计算出的累加相关系数会很小；过了 DD 点后，对应地层相同，累加相关系数从此点开始增大，在累加相关系数曲线上将会出现一个明显的台阶(图 9-5b)，这个台阶指示的位置即为 No1 井内断点的位置，它在 No2 井内对应的深度为 $DD1$。同理，在图 9-5c的位置时，累加相关系数曲线会出现一个相反的台阶，它也指示了 No1 井内断点的位置，该位置在 No2 井内对应的深度为 $DD2$。$DD1$ 与 $DD2$ 的差即为断层的落差。

根据 N 条累加相关系数曲线的特点，分析剖面上的地质现象，可更有效地进行地层对比。在 No1 和 No2 井对比结果图(图 9-6)的累加相关系数曲线上看出，在序号为 70 处出现台阶。但是我们计算累加相关系数的初始长度为 5，因此，断点位置应在序号 75 处，即 No1 井的 1 650 m 深处，也就是 No1 井缺失了 No2 井深度为 1 730 m 以上的地层。

【例 3】抑制随机干扰。

利用不同的滑动平均方程，消除或抑制季候泥厚度数据序列中的随机干扰，绘制新数据的曲线(图 9-7)。与原始数据曲线相比，消除或抑制随机干扰后的曲线更光滑。

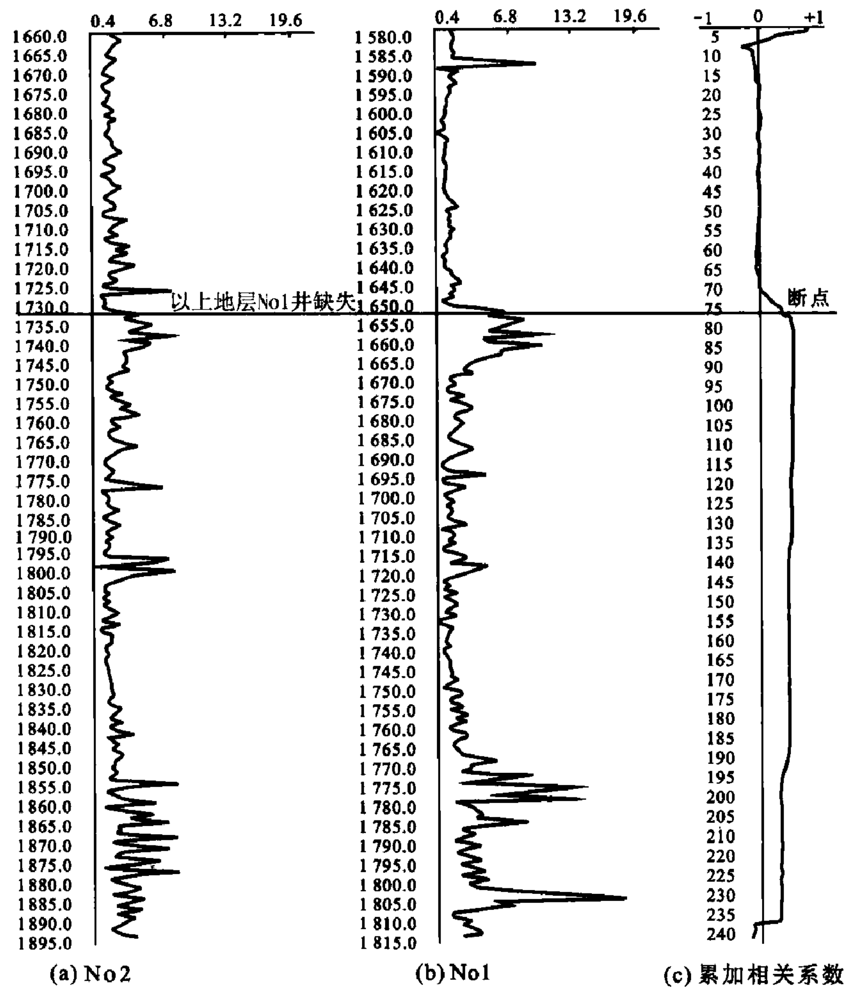
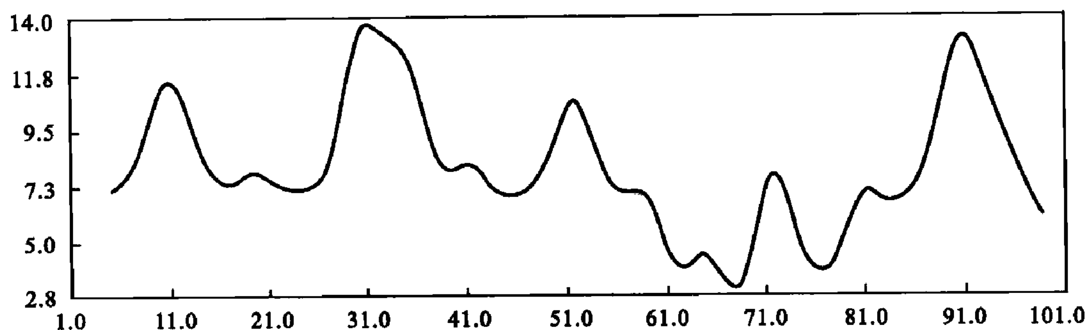
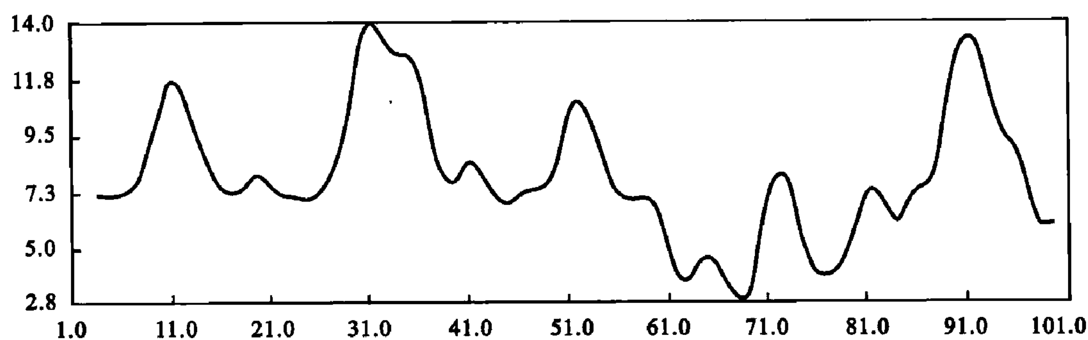


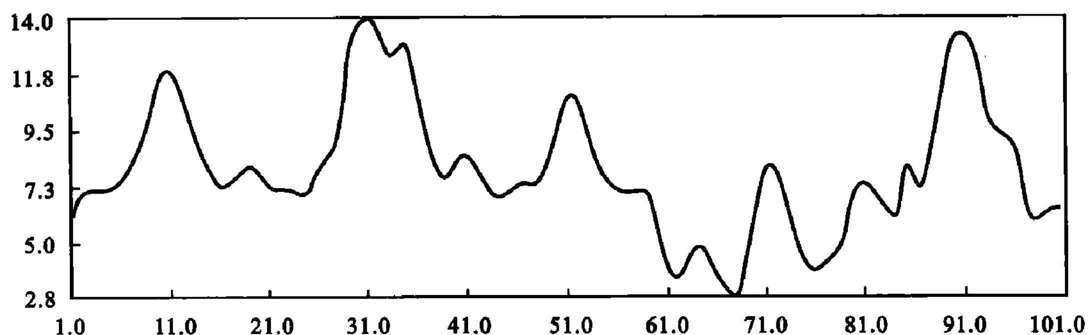
图 9-6 对比结果图



七项方程滑动平均后的曲线



五项方程滑动平均后的曲线



原始数据曲线

图 9-7　滑动平均后的曲线图

思考与练习

1. 什么是有序数据序列分析?
2. 什么是自相关、互相关分析?
3. 地质学中的哪一类问题属于相关分析的研究对象?
4. 试述滑动平均法抑制随机干扰的计算过程。
5. 试举例说明有序数列分析在石油及天然气地质中的应用。

第十章 油气资源量及含油气有利地带的预测

油气资源评价是贯穿整个油气勘探开发过程中的一项工作，而油气资源量及含油气有利地带的预测则是油气资源评价的重要任务之一。

20 世纪 70 年代以来，计算机技术的迅猛发展及其在油气勘探开发中的普遍应用，使数学地质方法的应用领域更加广泛，进而极大地促进了油气资源评价工作的进展，既提高了评价方法的科学性，又增强了评价结论的可靠性。在不同的油气勘探开发阶段，油气资源评价的任务、依据的理论基础也不一样，因此，也就有了不同的油气资源评价方法。据不完全统计，目前国内外用于油气资源评价的方法有百余种，其中大多数方法属于油气资源量的预测方法，而预测探区含油气有利地带的方法较少，这是目前油气资源评价工作中的一个薄弱环节。研究含油气有利地带的预测方法，不仅有利于提高油气资源评价结论的可靠性，更重要的是可用于指导探区当前的勘探工作。

在回归分析、聚类分析、趋势面分析、判别分析、蒙特卡罗法和模糊数学方法（第十一章）的应用实例中，扼要地介绍了几种预测油气资源量和含油气有利地带的方法。其中有的方法既可以预测资源量，也可以预测含油气有利地带，这取决于方法中所取地质变量的内涵。在此，介绍预测油气资源量的 Weng 旋回模型法、油田规模序列法、历史趋势外推法、地质圈闭的模糊集合综合评价法和多种信息叠合评价法。

§1 油气资源量预测

一、Weng 旋回模型法

1. Weng 旋回模型

若干个互相联系的事物或意识构成的一个整体称做一个体系，记为 Q。如果 Q 具有从兴起到衰亡的全过程，则称这一全过程为一个生命旋回，并称生命旋回中截止时间为 t 时 Q 的输出量为生命量。如果 Q 在 $t\leqslant 0$ 时不存在，那么 Q 就是一个不连续的体系，记为：

$$\begin{cases}0, & t\leqslant 0\\ Q, & t>0\end{cases}$$

若 Q 的发展速度 $\mathrm{d}Q/\mathrm{d}t$ 正比于发展过程中 Q 的当前状态，即

$$\mathrm{d}Q/\mathrm{d}t = Q[(x/t)-1] \qquad (t>0) \tag{10-1}$$

式中 $[(x/t)-1]$——比例因子；

x——Q 发展进入顶峰期后所经历的发展阶段数。

由微分方程式(10-1)得：

$$Q = At^{x}\mathrm{e}^{-t} \qquad (t>0) \tag{10-2}$$

式(10-2)是我国著名科学家翁文波先生提出的一种对生命总量有限的体系进行描述和预测的模型，称其为 Weng 旋回模型。由此模型可以看出：Q 的兴衰正比于两个因子，其兴起正比于时间 t 的 x 次方，其衰亡正比于时间 t 的负指数函数。因此，Q 是时间 t 的函数，而 t 是时间间隔与体系特征系数 C 的比值，为此把式(10-2)改写为：

$$\begin{cases} Q_t = At^x e^{-t} \\ t = (T - T_0)/C \end{cases} \quad (t > 0) \tag{10-3}$$

式中　T_0——体系的起始时刻；

T——体系发展过程中的某时刻；

x——某个正整数；

C,A——体系特征和影响系数。

2. Weng 旋回模型的性质

(1) $dQ_t/dt = Axt^{x-1}e^{-t} - At^x e^{-t} = At^x e^{-t}(x/t-1)$，即

$$dQ_t/dt = Q_t(x/t-1)$$

因此，Q 的发展速度正比于发展过程中的当前状态，并且

① 当 $0<t<x$ 时，$dQ_t/dt>0$，表示 Q 处于兴起期；

② 当 $t=x$ 时，$dQ_t/dt=0$，表示 Q 发展达到高峰期；

③ 当 $t>x$ 时，$dQ_t/dt<0$，表示 Q 发展进入衰亡期。

(2) $dQ_t/dt^2 = Q_t t^{-2}[(t-x)^2 - x]$，故当 $t=x-x^{1/2}$ 或 $t=x+x^{1/2}$ 时，$dQ_t/dt^2=0$。

3. 生命量与生命总量

生命量是生命旋回中截止时间为 t 时 Q 的累计输出量，记为 $\sum_t Q_t$，即

$$\sum_t Q_t = \int_0^t Q_t dt = \int_0^t At^x e^{-t} dt \tag{10-4}$$

生命总量是生命旋回中截止时间 t 为 ∞ 时 Q 的累计输出量，记为 $\sum_\infty Q_t$，即

$$\sum_\infty Q_t = \int_0^\infty Q_t dt = \int_0^\infty At^x e^{-t} dt$$

当上式中的 x 为正整数，$t\to\infty$时，生命总量

$$\sum_\infty Q_t = A\Gamma(x+1) = Ax! \tag{10-5}$$

由式(10-4)，(10-5)得：

$$\sum_t Q_t / \sum_\infty Q_t = 1 - e^{-t}\sum_{i=0}^{x}(t^i/i!) \tag{10-6}$$

当 $t=0$ 时，式(10-6)等于 0，即体系截止时间为 0 时的生命量为 0；当 $t=\infty$时，式(10-6)等于 1，表明体系的生命总量是存在的，并且

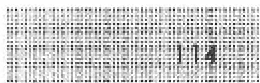
$$\sum_\infty Q_t = \sum_t Q_t / [1 - e^{-t}\sum_{i=0}^{x}(t^i/i!)] \tag{10-7}$$

因此，Weng 旋回模型是收敛的，它适合于对生命总量有限的体系进行描述和预测。

由上述可知，Q 的发展过程大体上可划分为加速上升阶段($0<t<x-x^{1/2}$)、一般上升阶段($x-x^{1/2}<t<x$)、稳定阶段($t=x\pm\Delta x$)、一般下降阶段($x<t<x+x^{1/2}$)以及缓慢下降阶段($x+x^{1/2}<t<\infty$)共五个发展阶段。

4. 油田产量及最终可采储量的预测

对于一个油田来说，从油田投产到产量枯竭是一个生命旋回，它是生命总量有限的体系，因此可以用式(10-3)和式(10-7)预测油田产量的未来变化和最终的可采储量。但是，油田的生命旋回将受开发措施和勘探等因素的影响，因此需要根据所预测的油田从投产年份开始的逐年实际产量，确定模型中的参数 x，C 和 A。



(1) 确定参数的原则。

设油田的逐年产量为 Q_i，模型预测的产量为 $Q_j = At^x e^{-t}(t>0)$。Q_j 与 Q_i 的相关程度越高，说明确定的参数 x,C 和 A 就越接近于油田的实际。也就是说，应该把 Q_j 与 Q_i 的相关程度达到最高作为确定模型中参数的基本原则。但是，预测模型中有三个参数，我们只能在上述原则下采用迭代的方法逐个确定每个参数。

(2) 确定参数的具体步骤。

① 确定参数 C 的初值。

根据油田逐年产量的变化，给出 C 的初始估计值。

② 确定参数 x 的估计值。

确定 C 的初值后，再进一步确定 x 的估计值。要求 Q_j 与 Q_i 最大相关，实际上等价于要求 $y_j = t^x e^{-t}$ 与 Q_i 最大相关。因此，令 $x_0 = 1,2,\cdots$，计算

$$\begin{cases} y_j = t^{x_0} e^{-t} \\ t = (T - T_0)/C \end{cases} \quad (t > 0) \quad (j = 1,2,\cdots,m) \tag{10-8}$$

的值，并计算 y_j 与逐年产量 $Q_i(i=1,2,\cdots,m)$ 的相关系数，当相关系数达到最大时对应的 x_0 即是给定 C 条件下 x 的一个估计值。

③ 确定参数 C 的最佳值。

在 x 估计值的基础上，用迭代的方法，取 $C=C+k\Delta C(k=1,2,\cdots;0<\Delta C<1)$，再利用式(10-8)计算 y_j 的值。当 y_j 与逐年产量 Q_i 的相关系数为最大时，确定 C 的最佳估计值。

确定参数 x,C 的过程可以反复进行，直到找到合适的值为止。

④ 确定参数 A。

设 Q_{t_i} 为逐年产量 Q_i 的预测值，记 $S = \sum_{i=1}^{m}(Q_i - Q_{t_i})^2 = \sum_{i=1}^{m}(Q_i - At_i^x e^{-t_i})^2$，令

$$dS/dA = 2\sum_{i=1}^{m}(Q_i - At_i^x e^{-t_i})(-t_i^x e^{-t_i}) = 0$$

整理得：

$$A = \sum_{i=1}^{m} Q_i t_i^x e^{-t_i} \Big/ \sum_{i=1}^{m}(t_i^x e^{-t_i})^2 \tag{10-9}$$

二、油田规模序列法

1. *油田规模序列的统计分布规律*

油田规模是指油田的最终可采储量。油田规模序列是指对含油区内已发现的油田的最终可采储量，按从大到小顺序排列成的数据序列。

假设 $Q_k(k=1,2,\cdots,n)$ 是某个含油区内第 k 个油田的规模，并且 $Q_1 \geqslant Q_2 \geqslant \cdots \geqslant Q_n$，那么国内外许多含油气区的统计资料表明，若以 $\ln Q_k$ 和 $\ln k$ 为纵、横坐标作散点图，那么各点大致分布在一条直线上(图 10-1)。

王幼梅根据我国含油气盆地实际资料，在直角坐标系中以油气田储量和规模序号作图(1977)，得到主要含油气盆地中油(气)田规模分布图(图 10-2)。

由图 10-2 可以看出，油田规模在一定范围内具有下列幂函数关系：

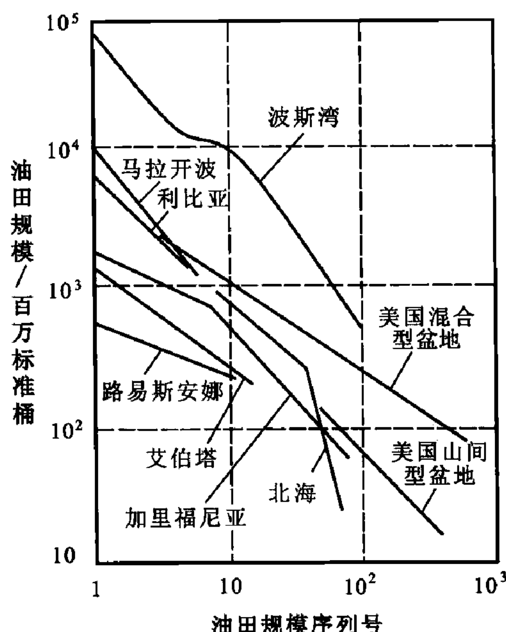


图 10-1　世界主要含油气区的油田规模

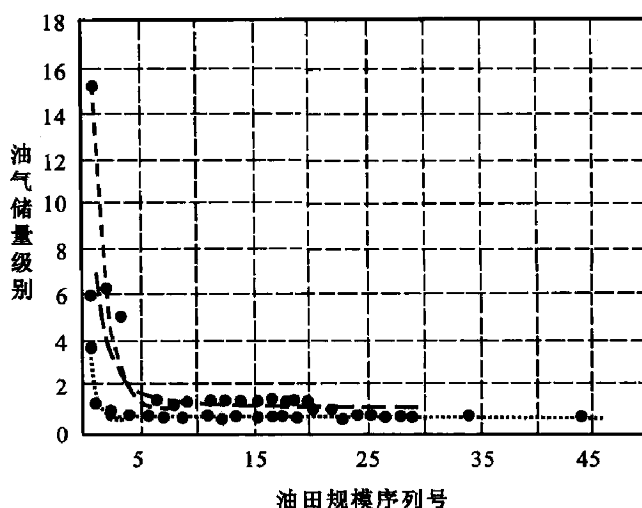


图 10-2　中国主要含油气盆地中油(气)规模分布图

$$Q_n = Q_1 n^{-k} \tag{10-10}$$

当 $k=1$ 时，

$$Q_n = Q_1/n \qquad (n = 1,2,\cdots)$$

对式(10-10)两边取对数，得：

$$\lg Q_n = \lg Q_1 - k\lg n$$

令

$$y = \lg Q_n , b = \lg Q_1 , x = \lg n$$

则有：

$$y = b - kx \tag{10-11}$$

式中　x——油田大小序号的对数；

k——直线的斜率；

b——盆地中已发现的最大油田储量的对数；

y——规模序号为 n 的油田储量的对数。

式(10-11)表示对数坐标系中的一组直线，这一结论与世界油田规模分布相一致。

有研究指出，斜率 k 可分为三类：$k<1$，$k=1$，$k>1$，分别代表分散型、过渡型和集中型盆地油田资源量的分布。

由上述可知，油田规模序列的统计分布规律是：在双对数坐标系中，油田规模的分布大致是条直线，并在一定范围内呈直线递减，直线段通常包含了油区 90%以上的储量，直线的急剧下倾，反映出油田规模随油田数目的增多而迅速下降。

2. 油田规模序列法

(1) 齐波夫定律。

美国学者齐波夫(Zipf)于 1949 年提出一个论断：把一组离散型随机变量的观测值按由大到小的顺序排列，其结果是最大的数值是第二大数值的两倍，是第三大数值的三倍，以此

类推。这一论断可表示为：

$$Q_1 = nQ_n \text{ 或 } Q_n/Q_m = m/n \quad (10\text{-}12)$$

式中 Q_1——序号为1的随机变量的观测值；

Q_n, Q_m——序号为 n, m 的随机变量的观测值；

$n, m(n \neq m)$——随机变量观测值的排列序号，可为任意正整数。

式(10-12)称为齐波夫定律。对式(10-12)两边取对数，整理得：

$$(\lg Q_m - \lg Q_n)/(\lg m - \lg n) = -1 \quad (10\text{-}13)$$

式(10-13)指出，在对数坐标系中作图时，直线的斜率等于－1。这一结论表明，过渡型盆地中油田规模序列服从齐波夫定律。

(2) 帕雷托定律。

帕雷托(Pareto)于1927年提出：

$$Q_m/Q_n = (n/m)^k \quad (10\text{-}14)$$

式中 Q_n, Q_m——序号为 n, m 的随机变量的观测值；

k——大于0的实数。

对式(10-14)两边求对数，整理得：

$$(\lg Q_m - \lg Q_n)/(\lg m - \lg n) = -k \quad (10\text{-}15)$$

式(10-15)表明，在双对数坐标系中，油田规模序列分布在斜率为 $-k$ 的直线上，这与世界主要含油气区油田规模统计规律相符合，即油田规模序列的分布服从帕雷托定律，而齐波夫定律是帕雷托定律中 $k=1$ 的特例。k 亦可叫做油田规模分布系数。

(3) 油田规模序列法。

一个含油气区内一组油田的石油储量是一组离散型随机变量，它们的分布规律服从帕雷托定律。因此，油田规模序列法是根据油气区内已发现的油田储量，利用帕雷托定律预测油区内尚未发现的油田储量(或资源量)以及全区石油总储量的一种油气资源估算方法。

3. 方法使用条件及注意事项

(1) 方法使用条件。

虽然世界上多数油气区的油田规模序列在一定程度上符合帕雷托定律，但到目前为止尚不能从油气田形成的地质理论上圆满解释油田规模序列的地质成因。事实表明，对于一个完整而独立的石油地质体系，油田规模序列法的预测效果较好。所谓一个完整而独立的石油地质体系是指该体系内的油气生成、运移、聚集以及其后的地质变迁都是在同一石油地质演化历史条件下发生的。简单来说，就是含油气区内油田(或油藏)应具有统一的地质成因。另外，本方法适用于含油气地区勘探的初期至晚期阶段。

(2) 注意事项。

利用油田规模序列法预测未发现油田时，要注意含油气区内油田的成油期。油田规模分布系数的不同，反映了油田规模序列的多样性。因此，当一个大的含油气地区具有多期成油过程时，就可能存在多个油田规模序列。在此情况下，应先把含油气区内的油田分类，然后按类应用油田规模序列法。

4. 油田规模序列法的计算过程

(1) 排序及选择推算点。

把油区内已发现的 t 个油田，按其储量由大到小排序，结果为 $Q_i(i=1,2,\cdots,t)$。选取

Q_1 作为推算点(预测油田储量规模的基准点),并假设 Q_1,$Q_i(i\neq1)$在油田规模预测模型中的序号分别为 m,n。

(2) 选择油田规模分布系数。

对于已发现的 t 个油田来说,虽知其规模序列的分布基本上服从帕雷托定律,但并不知道油田规模分布系数 k。在此情况下,先给出 k 的一个初值,进行多次油田规模序列的拟合计算,根据预测结果选择最终油田规模分布系数。由于 $k>0$,令 $k=-\tan\theta$,则 $\pi/2<\theta<\pi$。据统计,当 $2\pi/3\leqslant\theta\leqslant5\pi/6$ 时,可较快地选择到 k。记第 s 次选择的 k 为 k_s。

(3) 确定油田规模序列预测模型。

对于选定的 k_s,再进一步确定油田规模在预测模型中的序号 m 和 n,就可以给出一个相应的预测模型:

$$Q_i/Q_1=(m/n_i)^{k_s}$$

并以此预测含油气区油田规模序列。确定 m 和 n_i 的具体步骤如下:

① 令 $A_i=(Q_i/Q_1)^{1/k_s}(i=3,4,\cdots,t)$,则有

$$A_i=m/n_i,\qquad m=A_in_i$$

② 计算矩阵

$$\boldsymbol{B}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1t} \\ b_{21} & b_{22} & \cdots & b_{2t} \\ \vdots & \vdots & \vdots & \vdots \\ b_{\beta1} & b_{\beta2} & \cdots & b_{\beta t} \end{bmatrix}$$

矩阵 $\boldsymbol{B}$ 中的元素

$$b_{ji}=A_in_i\quad(i=1,2,\cdots,t;j=1,2,\cdots,\beta)$$

式中 n_i——使乘积 A_in_i 最大限度地接近矩阵 $\boldsymbol{B}$ 中行号为 j 的正整数。

计算各行的标准偏差

$$\sigma_j=\left[\frac{1}{t}\sum_{i=1}^{t}(b_{ji}-\bar{b}_j)^2\right]^{1/2}\qquad(j=1,2,\cdots,\beta)$$

$$\bar{b}_j=\frac{1}{t}\sum_{i=1}^{t}b_{ji}$$

当 $\sigma_m(\sigma_m\in\sigma_j,j=1,2,\cdots,\beta)$小于给定的误差 ε 时,$b_{mi}=A_in_i\approx m$,即有:

$$Q_i/Q_1\approx(m/n_i)^{k_s}$$

此式表明已在允许的误差内符合帕雷托定律,故可把矩阵的第 m 行作为含油气区油田规模序列的预测模型。

(4) 计算已发现油田的储量在油田规模序列预测模型中的序号。

已发现油田储量在预测模型中的序号

$$n_i=b_{mi}/A_i\qquad(i=1,2,\cdots,t)$$

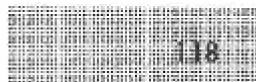
式中 n_i——已发现的第 i 个油田的储量 Q_i 在预测的油田规模序列中的序号;

b_{mi}——矩阵 $\boldsymbol{B}$ 中第 m 行第 j 列上的元素。

(5) 预测的最大油田储量。

预测的最大油田储量

$$\hat{Q}_{\max}=Q_in_i^{k_s}\qquad(i=1,2,\cdots,t)$$



即预测的最大油田储量为任一个已发现油田的储量 Q_i 乘以预测序号 n_i 的 k_s 次幂。在已发现的所有油田储量都可靠的情况下，取

$$\hat{Q}_{max} = \frac{1}{t}\sum_{i=1}^{t} Q_i n_i^{k_s}$$

为预测的最大油田储量。

(6) 计算油区内预测的油田规模序列。

设 Q_{min} 是最小经济油田的储量，那么预测的油田规模序列

$$\hat{Q}_j = \hat{Q}_{max} j^{-k_s} \qquad (j = 1,2,\cdots,p)$$

(7) 预测油田规模序列的循环计算。

在[$2\pi/3,5\pi/6$]内给 θ 一个增量 $\Delta\theta$，令 $k_s=\tan(\theta+\Delta\theta)$，重复上述(3)，(4)，(5)，(6)步的计算，得到下一个预测油田规模序列。若计算了 s 个预测油田规模序列，那么选择标准偏差

$$\sigma_r = [\frac{1}{t}\sum_{i=1}^{t}(Q_i - \hat{Q}_{r_i})^2]^{1/2} \qquad (r = 1,2,\cdots,s)$$

中最小的一个所对应的预测油田规模序列作为最终的油田规模序列模型。

(8) 油区石油总量。

由最终选定的油田规模序列计算油区石油总量

$$Q_z = \sum_{j=1}^{p} \hat{Q}_j$$

这仅是一个计算结果，是否合理还应与熟悉油区地质情况的研究人员商讨。

5. 油田规模序列法的讨论性小结

(1) 油田规模序列的理论预测模型——帕雷托定律。

近年来，国内很多研究和生产单位用齐波夫定律预测油田和金属矿床的储量，有时效果不佳，原因在于有的油田规模和矿床规模序列不服从齐波夫定律。实际上齐波夫定律仅是帕雷托定律中 $k=1$ 时的特例，因此应把帕雷托定律作为矿藏规模序列的理论预测模型。

(2) 预测石油储量的可行方法。

目前虽不能从地质成因理论解释油田规模序列的分布规律，但国内外主要含油气区的油田规模序列却普遍服从或近似服从帕雷托定律。因此，油田规模序列法是一种预测含油气区尚未发现的石油储量的可行方法。

(3) 油田储量的可靠性。

对于已发现的一批油田，只有其中储量可靠的油田规模序列才服从或近似服从帕雷托定律，所以在利用油田规模序列法预测尚未发现的石油资源量时，注意不要使用有可能增加储量或油田储量参数有待验证的油田数据，否则就不能使用由此得到的油田规模序列的预测模型预测含油气区中的石油储量。

(4) 油田的完整性和独立性。

油田的合理划分是油田规模序列法实施中的一个重要环节。与油藏的含义一样，油田也应是一个完整而独立的含油单元，不能以人为划分的采油矿区作为含油单元。

(5) 独立的油田规模序列。

在双对数坐标中，如果已发现的油田规模序列呈折线状分布，那么在一定意义上表明含

油气区存在不止一个相互独立的油田规模序列，此时应把已发现的油田分解成相互独立的油田规模序列。

(6) 预测结果的多解性。

油田规模序列的预测模型随着 k 的取值而变化，由此决定了预测结果的多解性。因此，在系数 k 不清楚的情况下，应多次变换 k 值，得到一组油田规模序列的预测模型，从中选出一个使已发现油田储量与对应预测值偏差最小的预测模型，并与熟悉含油区地质情况的地质工作者一起分析预测结果，确定符合目前实际的预测结果。

三、历史趋势外推法

1. 方法的基本思想

历史趋势外推法(经验外推法、历史状态法、历史趋势法、特性曲线法)是根据探区内已往的油气勘探工作量(如勘探钻井进尺、勘探井数或勘探时间等)以及与其相应的油气发现量，利用统计分析方法预测探区未来油气发现量的方法。该方法由 Davis(1958)提出，后来美国学者 Zopp(1962)，Hubbert(1967)等对此方法有所发展。该方法包括进尺发现率法、探井发现率法和年发现率法等。这些方法的基本思想是根据勘探工作量与油气发现量确定油气发现率曲线(图10-3)，并用该曲线预测未来的油气发现量。

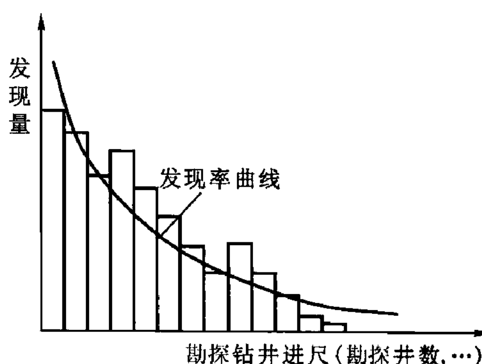


图 10-3　油气发现率曲线示意图

假如已往的勘探趋势和成功率能持续下去，则可将这类历史性资料拟合成曲线进行外推，以此确定未来的资源发现量。这种方法主要适用于勘探成熟区。在没有重大意外事件打乱探区发现率总趋势的情况下，多年沿用这种方法预测探区油气发现量是可行的。

2. 下降曲线模型

美国地质调查局主要采用指数下降曲线和双曲线下降曲线两个模型，即

$$\begin{cases} y = a\mathrm{e}^{bx} \\ y = ax^{-b} \end{cases} \tag{10-16}$$

式中　x——勘探工作量(钻井进尺、探井井数或时间等)；

y——油气体积。

根据历史资料，利用回归分析方法求出式(10-16)中参数 a 和 b 的估计值后，就可用其预测未来油气发现量。

3. 累计曲线模型

假设勘探阶段工作量为 d，勘探阶段数为 x，那么指数下降曲线可改写为：

$$y = a\mathrm{e}^{bdx}$$

对上式积分，得累计曲线模型

$$y = [a\mathrm{e}^{bd(n+m+1)} - a\mathrm{e}^{bd(n+1)}]/bd \tag{10-17}$$

式中　n——已进行过的勘探阶段数；

m——未来增加的勘探阶段数；

a,b——指数模型中的参数；

y——未来 m 个勘探阶段增加的待发现累计资源量。

设已进行过的 n 个勘探阶段增加的累计资源量为 v，那么 $n+m$ 个勘探阶段增加的总累计资源量

$$w = v + y \tag{10-18}$$

由下降曲线模型的性质可知，当 $m \to \infty$ 时，式(10-17)的极限就是最终待发现累计资源量

$$u = -ae^{bd(n+1)}/bd \tag{10-19}$$

最终总累计资源量

$$Q = v + u \tag{10-20}$$

经济效益是衡量油气勘探成效的一个重要指标。因此，应该根据历史趋势外推法预测的未来油气资源量的经济效益决定是否进入下一个勘探阶段。

§2 含油气有利地带预测

一、地质圈闭的模糊集合综合评价法

经过勘探发现了一批地质圈闭后，人们最关心的问题是这些圈闭中哪些含油，哪些可能含油，哪些不含油，这也就是通常所说的地质圈闭的含油性评价问题。这里给出的含油性概念是模糊的，因为不同含油性的过渡没有明显的界线。对地质圈闭的评价经常是上述模糊概念组成的模糊集合。地质学中的一些模糊概念，适于用模糊数学方法处理。

应用模糊数学方法对地质圈闭进行综合评价时，应考虑以下问题：

① 与地质圈闭含油性有关的多个控制油气形成的地质因素之间，可能存在多层次的结构关系。例如，地质圈闭的含油性通常取决于生油条件、储油条件和盖层条件等，这些基本条件又由在当时的勘探程度下可能取得的若干个次一级的地质因素构成，如生油条件可能与地质圈闭所处的生油条件分区、生油岩厚度、生油岩的地球化学指标等因素有关。

② 各种地质因素对综合评价所起的作用难以估计，多凭经验给出，常用权重分配表示。

③ 模糊集合综合评价中的矩阵合成运算有多种算子，仅用一种算子将会丢失很多信息，从而使评价结果过于单调，甚至难于鉴别地质圈闭含油性的优劣。

1. 地质圈闭综合评价的基本思想

探区经过地震勘探发现了一批地质圈闭，钻探前要根据圈闭的含油性对圈闭进行排队，从中选择含油性好的圈闭进行钻探。

评价地质圈闭含油性时，若用了 n 项地质因素，则构成 n 项因素集合，记为：

$$U = \{U_1, U_2, \cdots, U_n\}$$

式中 $U_i(i=1,2,\cdots,n)$是集合 U 的元素或子集。当 U_i 是 U 的子集时，它可由 n_i 项元素或次一级子集组成，即

$$U_i = \{U_{i1}, U_{i2}, \cdots, U_{in_i}\}$$

预想把圈闭的含油性分为 m 个级别，则可设评价集合为：

$$V = \{V_1, V_2, \cdots, V_m\}$$

考虑地质因素在评价地质圈闭含油性时的作用，假设因素集合 U 的权重分配为：

$$A = \{A_1, A_2, \cdots, A_n\}$$

式中 $A_i(i=1,2,\cdots,n)$是集合 A 的元素或子集。当 A_i 是 A 的子集时，它可由 n_i 项元素或

次一级子集组成，即

$$A_i = \{A_{i1}, A_{i2}, \cdots, A_{in_i}\}$$

这里要求 $\sum_{i=1}^{n} A_i = 1$，$\sum_{j=1}^{n_i} A_{ij} = 1$ 。

从 U 到 V 的一个模糊映射叫做单因素评价，当 $U_i \in U$ 时，有 $R(U_i) = (r_{i1}, r_{i2}, \cdots, r_{im})$。

对地质圈闭含油性进行综合评价时，若地质因素既有定量数据，又有相对关系或定性描述时，一律采用相对评语表示子集 $R(U_i)$。对于定量数据，可通过等级变换转化为相对评语。

当评价集合分为好、中、差三个级别时，可按表 10-1 中的评语级别表示子集 $R(U_i)$。与此类似，当评价集合分为好、较好、中等、较差、差五个级别时，可按表 10-2 中的评语级别表示子集 $R(U_i)$。

表 10-1 三个级别的评语表

评语 \ 级别	−1	0	1
好	0	0.2	0.8
中	0.25	0.5	0.25
差	0.8	0.2	0

表 10-2 五个级别的评语表

评语 \ 级别	−2	−1	0	1	2
好	0	0	0	0.2	0.8
较好	0	0	0.2	0.6	0.2
中等	0	0.25	0.5	0.25	0
较差	0.2	0.6	0.2	0	0
差	0.8	0.2	0	0	0

如果 U_i 是 U 的元素，则可由 n 个模糊映射 $R(U_i)$ 组成综合评价变换矩阵

$$\boldsymbol{R} = (r_{ij})_{n\times m} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}$$

如果 U_i 是 U 的子集，则可由 n_i 个模糊映射 $R(U_{ij}) = (r_{ij}^{(1)}, r_{ij}^{(2)}, \cdots, r_{ij}^{(m)})$ 组成单项地质因素的综合评价变换矩阵

$$\boldsymbol{R}_i = (r_{ij}^{(k)})_{n_i\times m} = \begin{pmatrix} r_{i1}^{(1)} & r_{i1}^{(2)} & \cdots & r_{i1}^{(m)} \\ r_{i2}^{(1)} & r_{i2}^{(2)} & \cdots & r_{i2}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ r_{in_i}^{(1)} & r_{in_i}^{(2)} & \cdots & r_{in_i}^{(m)} \end{pmatrix}$$

如果 U_i 是 U 的元素，那么地质因素的权重分配 A 与综合评价变换矩阵 $\boldsymbol{R}_h$ 的合成 $\boldsymbol{B}_h$ 称为第 h 个地质圈闭的综合评价，即

$$\boldsymbol{B}_h = A \circ \boldsymbol{R}_h \qquad (h = 1,2,\cdots,p) \tag{10-21}$$

如果 U_i 是 U 的子集，而 U_{ij} 是 U_i 的元素，那么首先计算次一级地质因素的综合评价 $\boldsymbol{B}_{hi}$，即

$$\boldsymbol{B}_{hi} = A_i \circ \boldsymbol{R}_{hi} \qquad (h = 1,2,\cdots,p; i = 1,2,\cdots,n) \tag{10-22}$$

式中 i——地质因素的编号；

h——地质圈闭的编号；

$\circ$——矩阵合成算子。

次一级地质因素综合评价的计算结果应作为上一级综合评价变换矩阵的一行。

最后，可用下式求得每个地质圈闭含油性的综合评价值 D_h：

$$D_h = \boldsymbol{B}_h \boldsymbol{C}' \qquad (h = 1,2,\cdots,p) \tag{10-23}$$

式中 $\boldsymbol{C}'$——等级矩阵的转置矩阵。

当评价集合为好、中、差三个级别时，可取等级矩阵为：

$$\boldsymbol{C} = (-1,0,1)$$

求出 p 个地质圈闭含油性的综合评价值 D_h 后，可按 D_h 的大小将地质圈闭按含油性的相对好坏进行排序，并按该次序钻探探区中已发现的地质圈闭。

2. 矩阵合成算子

地质因素的权重分配 A 与综合评价矩阵 $\boldsymbol{R}$ 的合成称为地质圈闭的综合评价 $\boldsymbol{B}$，即

$$A \circ \boldsymbol{R} = \boldsymbol{B} = (b_1, b_2, \cdots, b_m)$$

式中的 $\circ$ 为矩阵合成算子，常用的合成算子有以下 4 种；

（1）取小取大运算。

合成矩阵 $\boldsymbol{B}$ 中的元素 b_j 的计算方法为：

$$b_j = \bigvee_{i=1}^{n} (a_i \wedge r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{10-24}$$

式中 $\wedge$——在 a_i，r_{ij} 中取最小值；

$\vee$——在 n 个最小值中取一个最大值。

（2）乘积取大运算。

合成矩阵 $\boldsymbol{B}$ 中的元素 b_j 的计算方法为：

$$b_j = \bigvee_{i=1}^{n} (a_i \cdot r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{10-25}$$

（3）取小求和运算。

合成矩阵 $\boldsymbol{B}$ 中的元素 b_j 的计算方法为：

$$b_j = \bigoplus_{i=1}^{n} (a_i \wedge r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{10-26}$$

式中 $\oplus$——n 个最小值求和。

（4）乘积求和运算。

合成矩阵 $\boldsymbol{B}$ 中的元素 b_j 的计算方法为：

$$b_j = \bigoplus_{i=1}^{n} (a_i \cdot r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{10-27}$$

例如：

$$A = \{0.6, 0.3, 0.2\}, \qquad \boldsymbol{R} = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.4 & 0.2 & 0.4 \\ 0.1 & 0.6 & 0.3 \end{bmatrix}$$

按取小取大运算、乘积取大运算、取小求和运算、乘积求和运算的结果分别为：(0.3,0.3,0.5)，(0.12,0.18,0.30)，(0.6,0.7,1.0)，(0.26,0.36,0.48)。

二、多种信息叠合评价法

多种信息叠合评价法是对探区已有资料进行综合处理的一种方法。这种综合处理方法的目的是想得到与含油气有利地带关系更为密切的综合信息，为制定探区的勘探方案提供依据。

含油气有利地带的预测，不应局限于钻探井位的选择，而且还应该包括全国范围内的有利含油气盆地、盆地内含油气有利地质坳陷的挑选、坳陷内含油气有利凹陷或凸起的确定、凹陷或凸起内有利圈闭的排队以及圈闭上最佳井位的确定等一系列预测工作。下面介绍预测含油气有利地带的多种信息叠合评价法。

1. *多种信息叠合评价法的基本思想*

多种信息叠合评价法的基本思想是：先把控制油气形成的各种单一地质因素作为基础地质信息，将其绘制成基础地质信息图，再由概念不同的基础地质信息图叠加得到组合地质信息图，最后将组合地质信息图叠加生成综合地质信息图。在该图的基础上，进行综合地质解释，预测含油气有利地带。

综上所述，多种信息叠合评价法的基本思想可以概括为“图叠图得新图”。新图所包含的信息是基础地质信息经逐级叠加而得到的综合性信息。

2. *多种信息叠合评价法的实施步骤*

(1) 地质信息的归类与分级。

整理收集到的地质信息，形成归类合理、层次分明的信息体系，通常分为两个层次，即基础地质信息和组合地质信息。同类的基础地质信息构成组合地质信息，组合地质信息叠加构成综合地质信息。例如在某探区收集到生油岩厚度、生油岩有机碳含量、储集层厚度、储集层孔隙度、储集层渗透率、盖层厚度、局部构造特征等共七项与油气形成有关的地质信息，归类后得到如表 10-3 所示的层次关系信息体系。

表 10-3　地质信息体系表

基础地质信息	组合地质信息
生油岩厚度、生油岩有机碳含量	生油条件
储集层厚度、储集层孔隙度、储集层渗透率	储集条件
盖层厚度	盖层条件
局部构造特征（构造面积、闭合高度）	构造条件

(2) 基础地质信息的平面插值。

在基础地质信息分布稀疏离散的情况下，要对基础地质信息进行平面插值处理。插值的方法有很多种，可以根据基础地质信息的实际情况、选择合适的插值方法。

(3) 生成基础地质信息图。

在地质信息归类分级和插值处理的基础上，生成统一比例尺的基础地质信息图。

(4) 生成组合地质信息图。

把同类的基础地质信息图叠加，形成组合地质信息图。

(5) 生成综合地质信息图。

把不同的组合地质信息图叠加，形成综合地质信息图。

3. 地质信息叠合方法

多种信息叠合评价法的实施步骤，是从“图叠图得新图”角度讲的。利用计算机进行地质信息叠合时，是把地质信息值按某种约定的算法叠加，得到新的地质信息。下面介绍基础地质信息和组合地质信息的叠合方法。

(1) 地质信息的叠合前处理。

① 地质信息的正规化。

为了保持各种地质信息在叠合中的等价性及可加性，一般采用极差正规化方法，将各种基础地质信息变换到[0,1]区间内。

② 地质信息的权。

在进行叠合之前，应当根据各种基础地质信息或组合地质信息对油气形成所起的作用，赋以合理的权，以体现它们所起作用的大小。

(2) 基础地质信息叠合方法。

基础地质信息叠合方法是指由基础地质信息叠合生成组合地质信息的方法。

① 乘积叠合。

这种叠合方法是把平面上同一坐标点的 m 种基础地质信息值进行加权连乘，得到该点的组合地质信息值，即

$$z_j = \prod_{i=1}^{m} \lambda_{ji} z_{ji} \qquad (j = 1,2,\cdots,w) \tag{10-28}$$

式中 z_{ji}——第 j 种组合地质信息值中的第 i 种基础地质信息值；

z_j——m 种基础地质信息值叠合而成的第 j 种组合地质信息值；

w——组合地质信息种类数；

λ_{ji}——第 j 种组合地质信息中第 i 种基础地质信息值的权。

在乘积叠合图基础上对探区进行分带评价时，分带区间不是等间距的，区间间隔值为：

$$H_r = [(r-1)/k]^m \qquad (r = 1,2,\cdots,k+1) \tag{10-29}$$

式中 H_r——第 r 个分带区间间隔值；

k——分带区间总数；

m——叠合时基础地质信息的个数。

② 累加叠合。

这种叠合方法是把平面上同一坐标点的 m 种基础地质信息值进行累加，得到该点的组合地质信息值，即

$$z_j = \sum_{i=1}^{m} \lambda_{ji} z_{ji} \qquad (j = 1,2,\cdots,w) \tag{10-30}$$

在累加叠合图基础上对探区进行分带评价时，分带区间是等间距的，区间间隔值为：

$$H_r = [(r-1)/k] \cdot m \qquad (r = 1,2,\cdots,k+1) \tag{10-31}$$

③ 取小叠合。

该方法是把平面上同一坐标点的 m 种基础地质信息值中的最小值作为叠合值，即

$$z_j = \min_{1 \leqslant i \leqslant m} (\lambda_{ji} z_{ji}) \qquad (j = 1,2,\cdots,w) \tag{10-32}$$

在取小叠合图基础上对探区进行分带评价时，分带区间是等间距的，区间间隔值为：

$$H_r = (r-1)/k \qquad (r = 1,2,\cdots,k+1) \tag{10-33}$$

(3) 组合地质信息叠合方法。

组合地质信息叠合是指把不同的组合地质信息叠合在一起，生成综合地质信息。如果信息体系分两个层次，按上述三种叠合方法组合，共有九种组合地质信息叠合方法。

① 双重乘积叠合。

双重乘积叠合法得到的综合地质信息值为：

$$z = \prod_{j=1}^{w} k_j z_j = \prod_{j=1}^{w} k_j \prod_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-34}$$

式中 z_j——第 j 种组合地质信息值；

k_j——第 j 种组合地质信息值的权。

② 乘积累加叠合。

乘积累加叠合法得到的综合地质信息值为：

$$z = \prod_{j=1}^{w} k_j z_j = \prod_{j=1}^{w} k_j \sum_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-35}$$

③ 乘积取小叠合。

乘积取小叠合法得到的综合地质信息值为：

$$z = \prod_{j=1}^{w} k_j z_j = \prod_{j=1}^{w} k_j [\min_{1 \leqslant i \leqslant m} (\lambda_{ji} z_{ji})] \tag{10-36}$$

④ 双重累加叠合。

双重累加叠合法得到的综合地质信息值为：

$$z = \sum_{j=1}^{w} k_j z_j = \sum_{j=1}^{w} k_j \sum_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-37}$$

⑤ 累加乘积叠合。

累加乘积叠合法得到的综合地质信息值为：

$$z = \sum_{j=1}^{w} k_j \prod_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-38}$$

⑥ 累加取小叠合。

累加取小叠合法得到的综合地质信息值为：

$$z = \sum_{j=1}^{w} k_j z_j = \sum_{j=1}^{w} k_j [\min_{1 \leqslant i \leqslant m} (\lambda_{ji} z_{ji})] \tag{10-39}$$

⑦ 双重取小叠合。

双重取小叠合法得到的综合地质信息值为：

$$z = \min_{1 \leqslant j \leqslant w} k_j [\min_{1 \leqslant i \leqslant m} (\lambda_{ji} z_{ji})] \tag{10-40}$$

⑧ 取小乘积叠合。

取小乘积叠合法得到的综合地质信息值为：

$$z = \min_{1 \leqslant j \leqslant w} k_j \prod_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-41}$$

⑨ 取小累加叠合。

取小累加叠合法得到的综合地质信息值为：

$$z = \min_{1 \leqslant j \leqslant w} k_j \sum_{i=1}^{m} \lambda_{ji} z_{ji} \tag{10-42}$$

上面介绍了多种信息叠合评价法的基本思想、叠合过程和叠合方法。但是，对不同地质信息进行叠合，获得一个新的综合地质信息，这个综合地质信息应有明确的地质含义，这是选择叠合方法时应注意的一个问题。

§3 应用实例

【例 1】前苏联罗马什金油田发现于 1948 年，其从 1952 年投产到 1979 年为止的逐年产量见表 10-4。

表 10-4 罗马什金油田(1952—1979)逐年产量

生产年份	年产量/($\times 10^4$ t)	生产年份	年产量/($\times 10^4$ t)	生产年份	年产量/($\times 10^4$ t)
1952	200	1962	5 000	1972	8 000
1953	300	1963	5 600	1973	8 000
1954	500	1964	6 040	1974	8 000
1955	1 000	1965	6 600	1975	8 000
1956	1 400	1966	6 800	1976	7 775
1957	1 900	1967	7 000	1977	7 500
1958	2 400	1968	7 600	1978	7 230
1959	3 050	1969	7 900	1979	6 800
1960	3 800	1970	8 150		
1961	4 400	1971	8 000		

根据 28 年的实际年产量，按照确定参数的原则，经计算得预测罗马什金油田产量变化的具体模型为

$$\begin{cases} Q_t = 6\ 002.24 t^3 e^{-t} \\ t = (T - 1\ 951)/6.78 \end{cases} \tag{10-43}$$

式中 T——预测年份。

利用式(10-43)预测的罗马什金油田逐年产量见表 10-5，逐年预测产量与逐年产量的相关系数 $r=0.998\ 4$。产量预测结果如图 10-4 所示。

表 10-5 罗马什金油田的预测的逐年产量

预测年份	预测产量/($\times 10^4$ t)	预测年份	预测产量/($\times 10^4$ t)	预测年份	预测产量/($\times 10^4$ t)
1952	16.62	1962	5 060.52	1972	8 056.05
1953	114.71	1963	5 668.98	1973	7 992.40
1954	334.06	1964	6 219.23	1974	7 880.20
1855	683.26	1965	6 702.47	1975	7 725.60
1956	1 151.48	1966	7 113.27	1976	7 534.66
1957	1 716.91	1967	7 449.04	1977	7 313.21
1958	2 352.51	1968	7 709.60	1978	7 066.82
1959	3 030.07	1969	7 896.74	1979	6 800.67
1960	3 722.67	1970	8 013.75	1980	6 519.53
1961	4 406.28	1971	8 065.10	1981	6 227.75

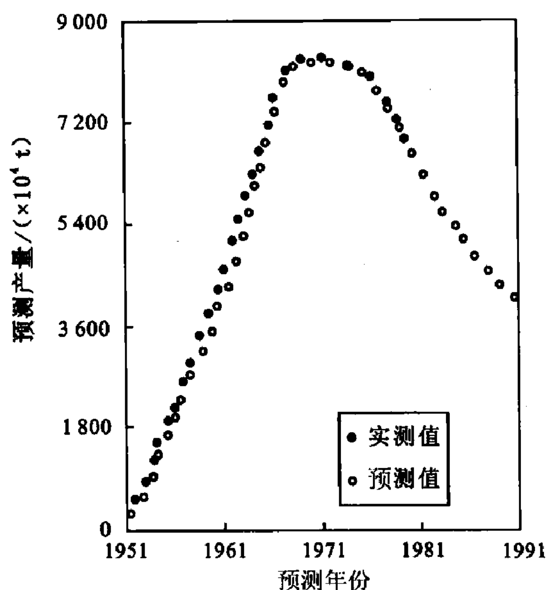


图 10-4 罗马什金油田产量预测图

把 $x=3$，$t=(1\ 979-1\ 951)/6.78$ 代入式(10-7)，预测罗马什金油田的最终可采储量为 $\sum_{\infty} Q_t = 24.988\ 2 \times 10^8 \mathrm{t}$ 。

【例 2】利用指数下降曲线模型预测美国密执安州北部 Niagaran 塔礁的将发现资源量、累计将发现资源量、累计总发现资源量。

表 10-6 中列出了由 Dan Gill 收集的 1968—1983 年 Niagaran 塔礁趋势带历史资料的外推统计表。

表 10-6 Niagaran 塔礁趋势历史资料的外推统计表

已钻探井数	发现资源量	累计发现量	今后钻探井数	指数下降曲线模型		
				将发现资源量	累计将发现资源量	累计总发现资源量
200	237	237	2 200	7.164 9	7.164 9	1 040.164 9
400	244	481	2 400	4.861 8	12.026 7	1 045.026 7
600	229	710	2 600	3.298 9	15.325 6	1 048.325 6
800	107	817	2 800	2.238 5	17.564 1	1 050.564 1
1 000	77	894	3 000	1.518 9	19.083 0	1 052.083 0
1 200	48	942	3 200	1.030 7	20.113 7	1 053.113 7
1 400	47	989	3 400	0.699 4	20.813 1	1 053.813 1
1 600	22	1 011	3 600	0.474 5	21.287 6	1 054.287 6
1 800	11	1 022	3 800	0.322 0	21.609 6	1 054.609 6
2 000	11	1 033	4 000	0.218 5	21.828 2	1 054.828 2

根据表 10-6 中已钻探井数和发现资源量数据进行回归分析，确定指数下降曲线模型 $y=ae^{bx}$ 中 $a=510.300\ 7$，$b=-0.001\ 939$，即 Niagaran 塔礁将发现资源量的指数下降曲线预测模型为：

$$y = 510.300\ 7e^{-0.001\ 939x} \tag{10-44}$$

式中 y——将发现资源量；

x——今后钻探井数。

预测结果见表 10-6 中将发现资源量、累计将发现资源量和累计总发现资源量。

【例 3】含油有利圈闭的选择。

根据地质调查和地震勘探，在某探区发现了 5 个地质圈闭(表 10-7)。

表 10-7　地质圈闭各地质因素的评语及数据

地质因素 \ 圈闭编号		1	2	3	4	5
生油条件		较　差	中　等	较　好	较　好	中　等
储油条件		中　等	较　好	中　等	中　等	较　差
该层条件		较　好	中　等	中　等	较　好	好
构造条件	面积/km²	10	20	10	15	20
	幅度/m	100	200	100	150	90
	断层条数	2	1	0	0	1

由于该探区的勘探程度较低，故生油条件、储油条件、盖层条件都是定性描述。构造条件由地震资料定量描述。根据现有资料，确定 5 个地质圈闭的先后勘探顺序。

(1) 构造条件的评语。

根据表 10-8 中规定的标准，将表 10-7 中定量描述的构造条件分为 5 个等级，并转换为评语描述，最后得到地质圈闭各地质因素的评语(表 10-9)。

表 10-8　各构造因素评语分级标准

地质因素 \ 评　语	差	较　差	中　等	较　好	好
构造面积/m²	<5	5～10	10～30	30～50	>50
构造幅度/m	<50	50～100	100～200	200～300	>300
断层条数	>2	1～2	1	0	0

表 10-9　地质圈闭各地质因素的评语

地质因素 \ 圈闭编号		1	2	3	4	5
生油条件		较　差	中　等	较　好	较　好	中　等
储油条件		中　等	较　好	中　等	中　等	较　差
该层条件		较　好	中　等	中　等	较　好	好
构造条件	面积/km²	较　差	中　等	较　差	中　等	中　等
	幅度/m	较　差	中　等	较　差	中　等	较　差
	断层条数	较　差	中　等	好	好	中　等

(2) 地质因素的权重分配。

经熟悉资料的人员分析，确定各个地质因素的权重分配如下：

$$
\text{综合评价}\begin{cases}
\text{生油条件}(0.25) \\
\text{储油条件}(0.25) \\
\text{盖层条件}(0.15) \\
\text{构造条件}(0.35)\begin{cases}\text{面积} & (0.4) \\ \text{幅度} & (0.3) \\ \text{断层条数} & (0.3)\end{cases}\text{子集}
\end{cases}
$$

即两级地质因素的权重分配为：

$$A = \{0.25, 0.25, 0.15, 0.35\}, \qquad A_4 = \{0.4, 0.3, 0.3\}$$

(3) 地质因素及评语集合。

选用的地质因素层次为生油条件 U_1、储油条件 U_2、盖层条件 U_3、构造条件 U_4，4 项地质因素构成因素集合 $U = \{U_1, U_2, U_3, U_4\}$，其中 U_1, U_2, U_3 是 U 的元素，U_4 是子集。子集 U_4 由构造面积、构造幅度、断层条数这 3 项次一级地质因素组成，即 $U_4 = \{U_{41}, U_{42}, U_{43}\}$。因各个地质因素和次一级构造因素均划分为 5 个级别，故构成 5 个级别的评语集合为 $V = \{V_1, V_2, \cdots, V_5\}$。

(4) 地质圈闭排队。

① 构造条件综合评价。

首先求次一级构造因素，即构造面积、构造幅度、断层条数的综合评价。按表 10-2 中的 5 个级别评语形成 $\boldsymbol{R}_{14}$，它是 1 号圈闭第 4 项地质因素的综合评价变换矩阵。若按乘积求和计算，1 号圈闭第 4 项地质因素的综合评价 $\boldsymbol{B}_{14}$ 为：

$$\boldsymbol{B}_{14} = A_4 \circ \boldsymbol{R}_{14} = \{0.4, 0.3, 0.3\} \circ \begin{bmatrix} 0.2 & 0.6 & 0.2 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 & 0 \end{bmatrix}$$

$$= (0.2, 0.6, 0.2, 0, 0)$$

按同样的方法计算，得 2,3,4,5 号圈闭的构造条件综合评价为：

$$\boldsymbol{B}_{24} = (0, \quad 0.25, \quad 0.5, \quad 0.25, \quad 0 \quad)$$

$$\boldsymbol{B}_{34} = (0.14, \quad 0.42, \quad 0.14, \quad 0.06, \quad 0.24)$$

$$\boldsymbol{B}_{44} = (0, \quad 0.175, \quad 0.35, \quad 0.235, \quad 0.24)$$

$$\boldsymbol{B}_{54} = (0.06, \quad 0.355, \quad 0.41, \quad 0.175, \quad 0 \quad)$$

② 地质圈闭综合评价。

进行地质圈闭综合评价时，上面计算得到的 $\boldsymbol{R}_{h4}$ 要作为综合评价矩阵 $\boldsymbol{R}_h$ 中的第 4 行。若仍按乘积求和计算，则 1 号圈闭的综合评价为：

$$\boldsymbol{B}_1 = A \circ \boldsymbol{R}_1 = (0.25, 0.25, 0.15, 0.35) \circ \begin{bmatrix} 0.2 & 0.6 & 0.2 & 0 \\ 0.2 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0.2 & 0.6 \\ 0.2 & 0.6 & 0.2 & 0 \end{bmatrix}$$

$$= (0.12, 0.4225, 0.275, 0.1525, 0.03)$$

按同样方法计算，得 2,3,4,5 号圈闭的综合评价为：

$$\boldsymbol{B}_2 = (0, \quad 0.1857, \quad 0.425, \quad 0.3375, \quad 0.07 \quad)$$

$$\boldsymbol{B}_3 = (0.049, \quad 0.247, \quad 0.299, \quad 0.271, \quad 0.134)$$

$$\boldsymbol{B}_4 = (0,\quad 0.123\,75,\quad 0.327\,5,\quad 0.384\,75,\quad 0.164)$$

$$\boldsymbol{B}_5 = (0.071,\quad 0.336\,75,\quad 0.318\,5,\quad 0.153\,75,\quad 0.12\ )$$

③ 圈闭排队。

按式(10-23)计算各地质圈闭的综合评价值 $D_h(h=1,2,\cdots,5)$：

$$D_1 = (0.12, 0.422\,5, 0.275, 0.152\,5, 0.03)\begin{pmatrix}-2\\-1\\0\\1\\2\end{pmatrix} = -0.45$$

按同样方法计算，得 2,3,4,5 号圈闭的综合评价值为：

$$D_2 = 0.29,\quad D_3 = 0.194,\quad D_4 = 0.589,\quad D_5 = -0.085$$

根据圈闭综合评价值，5 个地质圈闭编号排序为：4,2,3,5,1，其中 4 号地质圈闭的含油气地质条件最好，1 号地质圈闭的含油气地质条件最差，应首先对 4 号地质圈闭进一步勘探。

【例 4】含油气有利地带预测。

我国北方某沉积盆地是一个中新生代断陷盆地，沉积岩厚度超过 5 000 m，具备形成油气的基本地质条件。盆地中的 M 坳陷是最有远景的含油地区。为了提高勘探效益，评价该坳陷内各个地带的含油气地质条件，并在坳陷内寻找最有利的勘探地带是十分重要的。目前，勘探的主要目的层系是 k 系 p 组，钻探工作主要集中在坳陷的东部地区，全坳陷已基本完成地震勘探。因此，用多种信息叠合法进行评价时，是以地震勘探资料为主，结合钻井资料，对全坳陷进行有利勘探地带的预测。

评价中选择生油岩厚度、TTI 值、生油岩沉积相等共 10 项基础地质信息(表 10-10)，由它们叠合成生油条件、储油条件、构造条件、含油气状况和盖层条件共 5 项与油气形成有关的组合地质信息(表 10-10)，组合地质信息的再次叠加形成最终的综合地质信息，其叠合过程如图 10-5 所示。

表 10-10 M 坳陷的地质信息及分类标准

组合地质信息	基础地质信息	一 级	二 级	三 级
生油条件	生油岩厚度/m	<300	300～500	>500
	TTI 值	<8	>256	8～256
	生油岩沉积相	其 他	浅 湖	湖
储油条件	储集层厚度/m	<100	100～300	>300
	储集层沉积相	湖 沼	河流平原	浊积扇三角洲
构造条件	构造面积/km²	<20	20～35	>35
	构造幅度/m	<100	100～150	>150
	构造类型	断块、岩性	断鼻、半背斜、潜山	背 斜
含油气状况	油气产状	干 井	油气显示	油气流
盖层条件	盖层厚度/m	<1 200	>2 400	1 200～2 400

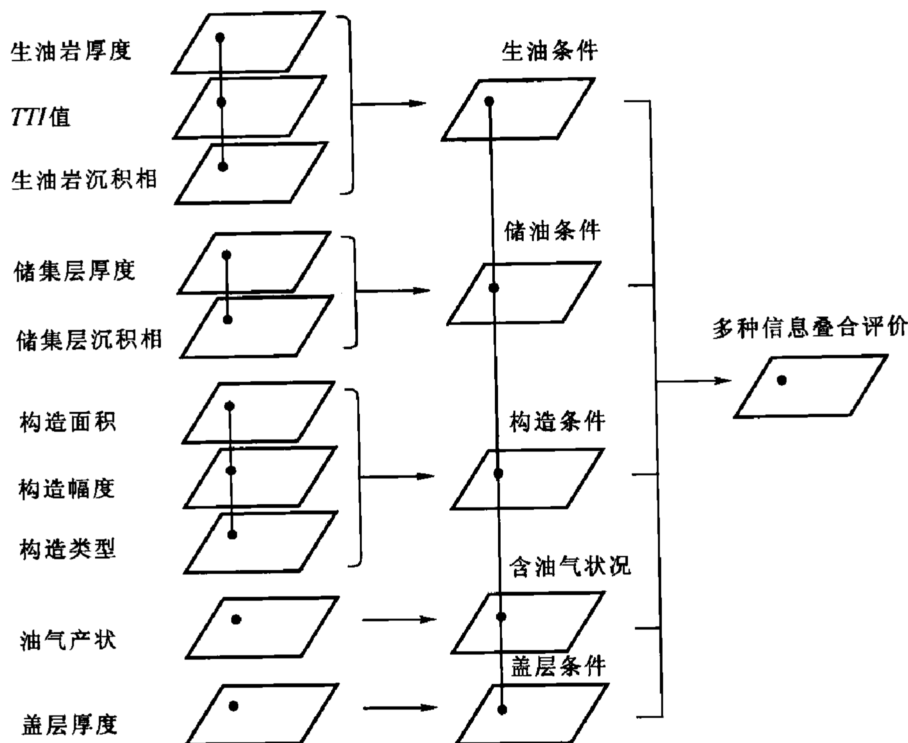


图 10-5 多种地质信息叠合过程示意图

根据地震、钻井资料在该坳陷内发现了 76 个局部构造。把所选的 10 项基础地质信息按统一标准划分为三个等级，给出分级标准与地质信息间的关系（表 10-10）。在三个级别中三级的含油气条件最好，一级最差。

为了便于计算，将基础地质信息按一、二、三级分别赋值为 1，2，3。按分级标准，形成坳陷 76 个局部构造的原始地质信息。原始地质信息的坐标均以原图的左下角为坐标原点，各点的坐标值是在原图上以 cm 为单位的实测值。

根据叠合选用的计算参数（表 10-11），由 10 项基础地质信息叠合得到生油条件、储油条件、构造条件、含油气状况和盖层条件 5 张组合地质信息评价图，最后由这 5 张组合地质信息评价图叠合得到综合地质信息评价图（图 10-6）。

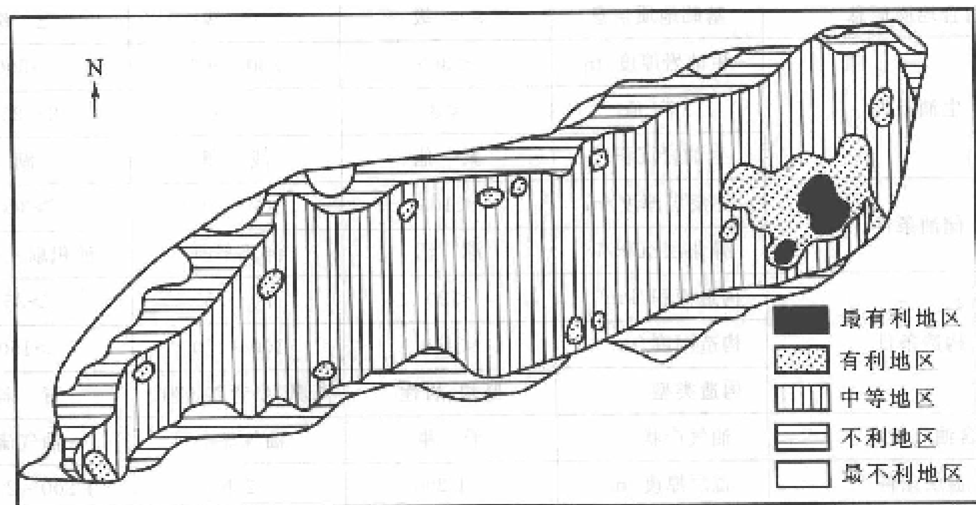


图 10-6 综合地质信息评价图

表 10-11 有关的叠合计算参数

基础地质信息名称	基础地质信息个数	插值方法	基础地质信息权	基础地质信息叠合方法	组合地质信息名称	组合地质信息中基础地质信息个数	组合地质信息权	组合叠合方法
生油岩厚度	76	圆　内	1	累加叠合	生油条件	5	1	累加叠合
TTI 值	76	圆　内	1					
生油岩沉积相	76	圆　内	1					
储集层厚度	76	圆　内	1	累加叠合	储油条件	2	1	
储集层沉积相	76	圆　内	1					
构造面积	76	圆　内	1	累加叠合	构造条件	3	1	
构造幅度	76	圆　内	1					
构造类型	76	圆　内	1					
油气产状	76	圆　内	1	累加叠合	含油气状况	1	1	
盖层厚度	76	圆　内	1	累加叠合	盖层条件	1	1	

这些评价图都划分为含油气最有利地区、有利地区、中等地区、不利地区和最不利地区共 5 个级别。在此基础上，选出 21 个含油气地质条件较好的局部构造，其中 4 个局部构造经钻探证实为含油构造。后期勘探证实，在图 10-6 东部的最有利地区，建成了具有一定规模油气产能的油田。

思考与练习

1. 何谓 Weng 旋回模型？它的描述和预测对象是什么？
2. 什么是生命旋回、生命量和生命总量？
3. 什么是齐波夫定律？什么是帕雷托定律？
4. 什么是油田规模？
5. 用齐波夫定律或帕雷托定律预测油田规模时应注意哪些条件？
6. 模糊数学方法为何适于对地质圈闭的含油气性进行综合评价？
7. 应用模糊数学方法对地质圈闭的含油气性进行综合评价时应注意哪些问题？
8. 多种信息叠合评价法的要点是什么？
9. 熟悉多种信息叠合评价法的实施步骤。
10. 信息叠合中应注意哪些问题？

第十一章　模糊数学方法及其应用

模糊数学是用数学方法研究和处理"模糊性"现象的数学。所谓的模糊性主要是指客观事物差异中间过渡界线的"不分明性"，如储层的含油气性、油田规模的大小、成油地质条件的优劣、圈闭的形态、岩石的颜色等。由于事物间差异的模糊性，因此描述它们特征的变量也是模糊的，即各变量的内部分级没有明显的界线。地质作用是复杂的，对其产生的地质现象有些可以采用定量的方法来度量，有些则不能用定量的数值来表达，而只能用客观模糊或主观模糊的准则进行推断或识别。1965 年美国控制论专家 Zadeh 提出模糊数学的概念后，模糊数学得到了迅速发展并逐渐应用到各个领域，在地质学中主要用于矿产资源评价，各种地质现象的分类、识别、决策和模拟。在此介绍油气勘探中常用的模糊聚类分析和模糊模型识别。

§1　模糊聚类分析

模糊聚类分析是在模糊相似矩阵的基础上，对分类对象进行定量分类的方法。它的主要内容包括数据的标准化、建立模糊相似矩阵和动态聚类三部分。

一、数据标准化

1. *原始数据*

假设有 n 个被分类对象，每个对象又有 m 个描述对象特性的变量，它们的观测值构成的原始数据矩阵为：

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

2. *极差正规化*

极差是变量观测值的最大值与最小值之差，即

$$\Delta x_j=\max_{1\leqslant i\leqslant n}x_{ij}-\min_{1\leqslant i\leqslant n}x_{ij}\qquad (j=1,2,\cdots,m)$$

极差正规化是用变量的每个观测值减去观测值的最小值后再除以极差。变换公式为：

$$x'_{ij}=(x_{ij}-\min_{1\leqslant i\leqslant n}x_{ij})/\Delta x_j\qquad (i=1,2,\cdots,n;j=1,2,\cdots,m)\quad (11\text{-}1)$$

式中　x'_{ij}——正规化后的数据；

x_{ij}——正规化前的数据(原始数据)；

$\min\limits_{1\leqslant i\leqslant n}x_{ij}$——第 j 个变量观测值的最小值。

由式(11-1)可知，对原始数据进行正规化处理后，新数据的最大值为 1，且 $x'_{ij}\geqslant 0$，即新数据分布在区间[0,1]内。

二、模糊相似矩阵

模糊相似矩阵是进行模糊聚类的基础。用于样品模糊聚类和变量模糊聚类的模糊相似

矩阵分别为 $\boldsymbol{R}$ 和 $\boldsymbol{R}'$。

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \qquad \boldsymbol{R}'=\begin{pmatrix} r'_{11} & r'_{12} & \cdots & r'_{1n} \\ r'_{21} & r'_{22} & \cdots & r'_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r'_{n1} & r'_{n2} & \cdots & r'_{nn} \end{pmatrix}$$

模糊相似矩阵中的元素是衡量分类对象之间模糊相似程度的指标，两种元素的计算方法类似。因此，下面仅介绍对样品进行模糊聚类时，建立模糊相似矩阵的常用方法，读者仿照这些方法容易写出对变量进行模糊聚类时，建立模糊相似矩阵的方法。为书写方便，在建立模糊相似矩阵的各种方法中，仍用 x_{ij} 表示正规化后的数据 x'_{ij}。

1. 相似系数法

(1) 数量积法。

$$r_{ij}=\begin{cases} 1 & i=j \\ \dfrac{1}{M}\sum\limits_{k=1}^{m} x_{ik}x_{jk} & i\neq j \end{cases} \qquad (i,j=1,2,\cdots,n) \tag{11-2}$$

式中 $M=\max\limits_{i\neq j}\left(\sum\limits_{k=1}^{m} x_{ik}x_{jk}\right)$。

显然，$|r_{ij}|\in[0,1]$。若 $r_{ij}<0$，令 $r'_{ij}=(r_{ij}+1)/2$，则 $r'_{ij}\in[0,1]$。r_{ij} 的大小指出分类对象 i 和 j 模糊相似程度。

(2) 最大最小法。

$$r_{ij}=\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})\Big/\sum_{k=1}^{m}(x_{ik}\vee x_{jk}) \qquad (i,j=1,2,\cdots,n) \tag{11-3}$$

式中 $\wedge$——在 x_{ik}，x_{jk} 两个元素中取一小值；

$\vee$——在 x_{ik}，x_{jk} 两个元素中取一大值。

(3) 算术平均最小法。

$$r_{ij}=2\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})\Big/\sum_{k=1}^{m}(x_{ik}+x_{jk}) \qquad (i,j=1,2,\cdots,n) \tag{11-4}$$

(4) 几何平均最小法。

$$r_{ij}=\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})\Big/\sum_{k=1}^{m}\sqrt{x_{ik}x_{jk}} \qquad (i,j=1,2,\cdots,n) \tag{11-5}$$

在式(11-3)，(11-4)，(11-5)中，要求 x_{ik}，x_{jk} 都大于 0，否则要对数据进行适当变换。显然 $r_{ij}\in[0,1]$。

(5) 夹角余弦和相关系数法。

参见第四章。

2. 距离系数法

(1) 绝对值倒数法。

$$r_{ij}=\begin{cases} 1 & i=j \\ M\Big/\sum\limits_{k=1}^{m}|x_{ik}-x_{jk}| & i\neq j \end{cases} \qquad (i,j=1,2,\cdots,n) \tag{11-6}$$

适当选取 M，使得 $0\leqslant r_{ij}\leqslant 1$。

(2) 切比雪夫距离法。

$$d_{ij}=\bigvee_{k=1}^{m}|x_{ik}-x_{jk}| \qquad (i,j=1,2,\cdots,n) \tag{11-7}$$

(3) 欧氏距离法。

参见第四章。

三、模糊聚类

1. 模糊等价矩阵

设论域 $U=\{x\}$ 是分类对象的一个集合。给定 U 上的一个模糊相似矩阵 $\boldsymbol{R}$，如果它满足自反性($r_{ij}=1$)、对称性($r_{ij}=r_{ji}$)和传递性($\boldsymbol{R}\circ\boldsymbol{R}\subseteq\boldsymbol{R}$)，则称 $\boldsymbol{R}$ 是 U 上的一个模糊等价矩阵。

这里的符号"∘"表示矩阵的合成运算，类似矩阵乘法运算，但要将元素的相乘改为取小值，相加改为取大值。

用相似系数法和距离系数法得到的模糊相似矩阵满足自反性和对称性，但不一定满足传递性。若模糊相似矩阵不满足传递性，则可先计算 $\boldsymbol{R}\circ\boldsymbol{R}$(记做 $\boldsymbol{R}^2$)，然后查看 $\boldsymbol{R}^2$ 是否满足传递性，若不满足，再经过 $\boldsymbol{R}^2\circ\boldsymbol{R}^2\cdots$ 运算，总可将 $\boldsymbol{R}$ 改造为满足传递性的模糊等价矩阵。

2. 模糊等价矩阵的 λ 截矩阵

设 $\boldsymbol{R}$ 是个模糊等价矩阵，对于任意的 $\lambda\in[0,1]$，称 $\boldsymbol{R}_\lambda$ 为 $\boldsymbol{R}$ 的 λ 截矩阵。$\boldsymbol{R}_\lambda$ 中的元素

$$r_{ij}^{(\lambda)}=\begin{cases}1 & r_{ij}\geqslant\lambda \\ 0 & r_{ij}<\lambda\end{cases} \qquad (i,j=1,2,\cdots,n)$$

式中 λ——阈值。

3. 动态聚类

由 $\boldsymbol{R}_\lambda$ 可知，当 $r_{ij}^{(\lambda)}=1$ 时，描述分类对象 i 与 j 的模糊变量的观测值最接近，即 i 与 j 之间的差异最小，应归为同一类，否则为不同的类。让 λ 由大到小变化，可形成动态聚类图。

四、最佳阈值 λ 的确定

在模糊等价关系的模糊聚类中，对于不同的 $\lambda\in[0,1]$，可得到不同的分类方案，从而形成一种动态聚类图，这对全面了解对象的分类情况是比较形象和直观的。但有的实际问题需要选择某个恰当的 λ，确定一个具体的分类方案，这就是确定阈值 λ 的问题。

1. 按实际需要确定

在动态聚类中，调整 λ 的值可以得到不同的分类，从中选择一种合适的分类方案。另外，也可由熟悉专业的专家确定阈值 λ，得到给定阈值 λ 水平下的分类。

2. 用 F 统计量确定 λ 的最佳值

设对应于 λ 的分类数为 r，第 j 类的对象数为 n_j，其样本记为 $\boldsymbol{x}_1^{(j)},\boldsymbol{x}_2^{(j)},\cdots,\boldsymbol{x}_{n_j}^{(j)}$，第 j 类的聚类中心为向量 $\bar{\boldsymbol{x}}^{(j)}=(\bar{x}_1^{(j)},\bar{x}_2^{(j)},\cdots,\bar{x}_m^{(j)})$。定义统计量

$$F=\frac{\sum_{j=1}^{r}n_j\|\bar{\boldsymbol{x}}^{(j)}-\bar{\boldsymbol{x}}\|^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j}\|\boldsymbol{x}_i^{(j)}-\bar{\boldsymbol{x}}^{(j)}\|^2/(n-r)}$$

式中 $\bar{\boldsymbol{x}}^{(j)}$——第 j 类中各变量的平均值，即 $\bar{\boldsymbol{x}}^{(j)}=(\bar{x}_1^{(j)},\bar{x}_2^{(j)},\cdots,\bar{x}_m^{(j)})$；

$\bar{x}_k^{(j)}$——第 j 类中第 k 个变量的平均值，即 $\bar{x}_k^{(j)}=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ik}^{(j)}(k=1,2,\cdots,m)$；

$\bar{\boldsymbol{x}}$——样品变量的平均值，即 $\bar{\boldsymbol{x}}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_m)$；

$\bar{x}_k$——第 k 个变量全部观测值的平均值，即 $\bar{x}_k=\frac{1}{n_j}\sum_{i=1}^{n}x_{ik}(k=1,2,\cdots,m)$；

$\|\bar{\boldsymbol{x}}^{(j)}-\bar{\boldsymbol{x}}\|^2$——$\bar{\boldsymbol{x}}^{(j)}$ 与 $\bar{\boldsymbol{x}}$ 距离的平方，即 $\|\bar{\boldsymbol{x}}^{(j)}-\bar{\boldsymbol{x}}\|^2=\sum_{k=1}^{m}(\bar{x}_k^{(j)}-\bar{x}_k)^2$；

$\|\boldsymbol{x}_i^{(j)}-\bar{\boldsymbol{x}}^{(j)}\|^2$——$\boldsymbol{x}_i^{(j)}$ 与 $\bar{\boldsymbol{x}}^{(j)}$ 距离的平方，即 $\|\boldsymbol{x}_i^{(j)}-\bar{\boldsymbol{x}}^{(j)}\|^2=\sum_{k=1}^{m}(x_{ik}^{(j)}-\bar{x}_k^{(j)})^2$。

F 服从第一自由度为 $r-1$，第二自由度为 $n-r$ 的 F 分布。它的分子是类与类之间的距离，分母是类内样品间的距离。F 越大，表明类与类之间样品的差异越大，即分类效果越好。

假设各类差异不明显，对于给定的检验水平 α，查 $F_\alpha(r-1,n-r)$ 分布表得临界值 F_α，若 $F<F_\alpha$，则认为各类之间有明显的差异，否则应重新选择阈值。

§2 模糊模型识别

模糊模型识别是已知论域 U 上有 n 个标准模糊模型，确定一个或一批待识别的模糊对象属于哪个标准模糊模型，即对待识别对象的归属做出判定。地质学领域中有很多属于模糊模型识别的问题，如在储层含油气性这一论域上，有油层、油气层、油水同层、气层、含水油层、干层等标准模糊模型，判断钻穿储层的含油气性就是一个对标准模糊模型的识别问题。为解决此类问题，必须构造一种衡量待识别的模糊对象与标准模糊模型之间模糊性程度的度量。在模糊模型识别中，常用的模糊性度量有隶属度和贴近度。

一、隶属度

1. 模糊向量及其内外积

若 $0\leqslant v_i\leqslant 1(i=1,2,\cdots,n)$，则称 $\boldsymbol{V}=(v_1,v_2,\cdots,v_n)$ 为模糊向量（模糊模型）。若 $\boldsymbol{X},\boldsymbol{Y}$ 是模糊向量，则称式(11-8)和式(11-9)分别为 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的内积和外积。

$$\boldsymbol{X}\circ\boldsymbol{Y}=\bigvee_{i=1}^{n}(x_i\wedge y_i) \tag{11-8}$$

$$\boldsymbol{X}*\boldsymbol{Y}=\bigwedge_{i=1}^{n}(x_i\vee y_i) \tag{11-9}$$

2. 模糊模型库

设 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$ 是论域 $U=\{x_1,x_2,\cdots,x_m\}$ 上的 n 个标准模糊模型，则称 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$ 为模糊模型库。

3. 隶属度

设 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$ 为模糊模型库，各模型的隶属函数为 $X_i(x)(i=1,2,\cdots,n)$，而 $\boldsymbol{x}^0=(x_1^0,x_2^0,\cdots,x_m^0)$ 为普通向量时，定义

$$X_i(\boldsymbol{x}^0)=\bigwedge_{j=1}^{m}[X_i(x_j^0)]\qquad(i=1,2,\cdots,n) \tag{11-10}$$

为 $\boldsymbol{x}^0$ 对 $\boldsymbol{X}_i$ 的隶属度。

隶属度适用于待识别对象是单因素的情形。用隶属度确定待识别对象的归属，关键是建立符合实际的隶属函数，但这是目前尚未完全解决的问题。我国学者汪培庄提出的随机集落影理论对于相当一部分模糊集的隶属函数的客观实在性给出了满意的解释，基于这一理论的模糊统计方法是确定一类模糊集隶属度的有效方法。目前确定隶属函数的方法有模

糊统计法、指派法、借用已有尺度法等。

4. 最大隶属度原则

原则Ⅰ：设论域$U=\{x_1,x_2,\cdots,x_m\}$上的模糊模型库为$\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$，若对任意$x\in U$，有$k\in(1,2,\cdots,n)$，使得

$$X_k(x_0)=\bigvee_{k=1}^{n}X_k(x) \tag{11-11}$$

成立，则认为x_0相对隶属于$\boldsymbol{X}_k$。

【例 1】学习成绩等级判断。

在学习成绩$U=\{0,100\}$这一论域上确定三个表示学习成绩的模糊模型$\boldsymbol{A}=$“优”，$\boldsymbol{B}=$“良”，$\boldsymbol{C}=$“中”。当一位学生的数学成绩为 88 分时，应该评为哪个等级？

解：用指派法建立论域$U=\{0,100\}$上$\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$的隶属函数：

$$A(x)=\begin{cases}0 & 0\leqslant x\leqslant 80\\(x-80)/10 & 80<x\leqslant 90\\1 & 90<x\leqslant 100\end{cases}$$

$$B(x)=\begin{cases}0 & 0\leqslant x\leqslant 70\\(x-70)/10 & 70<x\leqslant 80\\1 & 80<x\leqslant 85\\(95-x)/10 & 85<x\leqslant 95\\0 & 95<x\leqslant 100\end{cases}$$

$$C(x)=\begin{cases}0 & 0\leqslant x\leqslant 70\\(80-x)/10 & 70<x\leqslant 80\\1 & 80<x\leqslant 100\end{cases}$$

将$x=88$代入隶属函数计算，得$A(88)=0.8$，$B(88)=0.7$，$C(88)=0$。根据最大隶属度原则Ⅰ，该同学的成绩应评为优。

原则Ⅱ：设论域$U=\{x_1,x_2,\cdots,x_n\}$上有一个标准模型$\boldsymbol{A}$，待识别对象有$x_1,x_2,\cdots,x_m\in U$。如果有某个x_k满足

$$A(x_k)=\bigvee_{i=1}^{m}[A(x_i)] \tag{11-12}$$

则x_k与模型$\boldsymbol{A}$最相似。

【例 2】择优录用。

设三名学生英语学习成绩论域$U=\{x_1,x_2,x_3\}$，在U上确定一个模糊模型$\boldsymbol{A}=$“优”。若三名学生的成绩分别为$x_1=70$，$x_2=84$，$x_3=90$，那么，根据英语成绩哪位学生应优先留校任教。

解：分别将他们的英语学习成绩代入模糊模型隶属函数

$$A(x)=\begin{cases}0 & 0\leqslant x\leqslant 80\\(x-80)/10 & 80<x\leqslant 90\\1 & 90<x\leqslant 100\end{cases}$$

计算得到$A(70)=0$，$A(84)=0.4$，$A(90)=1$。根据最大隶属度原则Ⅱ，英语成绩为 90 分的学生应优先留任。

二、贴近度

1. 贴近度的一般定义

设 $\boldsymbol{A},\boldsymbol{B}$ 是论域 U 上的两个模糊模型，则称

$$\sigma_0(\boldsymbol{A},\boldsymbol{B}) = [\boldsymbol{A}\circ\boldsymbol{B} + (1 - \boldsymbol{A} * \boldsymbol{B})]/2 \qquad (11\text{-}13)$$

为 $\boldsymbol{B}$ 与 $\boldsymbol{A}$ 的贴近度，其中 $\boldsymbol{B}$ 是被识别对象。

贴近度是描述模糊模型之间彼此靠近程度的指标，由我国学者汪培庄提出。由于研究的问题不同，贴近度也有不同的定义形式。

贴近度适用于模糊模型库，待识别对象可以是模糊模型或普通模型，模糊模型库中的每个模糊模型和待识别对象均有 m 个相同的特性变量。设 $\boldsymbol{X}_i(i=1,2,\cdots,n)$ 和 $\boldsymbol{Y}$ 分别是论域 U 上模糊模型库中的模糊模型和待识别对象，若把 $\boldsymbol{X}_i,\boldsymbol{Y}$ 视为论域 U 上的点，那么，贴近度越大，$\boldsymbol{Y}$ 与 $\boldsymbol{X}_i$ 的距离就越小，即待识别对象与模糊模型的特性越相似。因此，可根据贴近度的相对大小对 $\boldsymbol{Y}$ 进行归类。选择某种计算贴近度的方法，计算 $\boldsymbol{Y}$ 与 $\boldsymbol{X}_i$ 的贴近度，记为 $\sigma_0(\boldsymbol{X}_i,\boldsymbol{Y})(i=1,2,\cdots,n)$，若其中的最大者为 $\sigma_0(\boldsymbol{X}_k,\boldsymbol{Y})$，则认为 $\boldsymbol{Y}$ 与 $\boldsymbol{X}_k$ 同类。

2. 择近原则

设论域 U 上的模糊模型库为 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$，$\boldsymbol{Y}$ 为 U 上的待识别对象，若存在 $k\in(1,2,\cdots,n)$，使得

$$\sigma_0(\boldsymbol{X}_k,\boldsymbol{Y}) = \bigvee_{j=1}^{n} \sigma_0(\boldsymbol{X}_j,\boldsymbol{Y}) \qquad (11\text{-}14)$$

则待识别对象归入 $\boldsymbol{X}_k$ 类。

3. 实用贴近度

在实际工作中常用的几个计算贴近度的公式如下：

$$\sigma_1(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}}) \triangleq \sum_{k=1}^{m}[\tilde{A}(x_k) \wedge \tilde{B}(x_k)]\Big/\sum_{k=1}^{m}[\tilde{A}(x_k) \vee \tilde{B}(x_k)]$$

$$\sigma_2(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}}) \triangleq 2\sum_{k=1}^{m}[\tilde{A}(x_k) \wedge \tilde{B}(x_k)]\Big/\sum_{k=1}^{m}[\tilde{A}(x_k) + \tilde{B}(x_k)]$$

$$\sigma_3(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}}) \triangleq 1 - \frac{1}{m}\sum_{k=1}^{m}|[\tilde{A}(x_k) - \tilde{B}(x_k)]|$$

§3 应用实例

【例 1】预测含油气有利地区。

为了寻找含油气有利的勘探地区，选择了 8 个生油指标和 5 个其他指标作为胜利油田 18 个洼陷的评价指标(表 11-1，据诸克军)，各指标的数据由专家组根据已有的勘探资料综合评分给定，满分为 10 分。已知利津洼陷是 18 个洼陷中勘探程度最高、含油气性最好的洼陷。

在对表 11-1 中数据进行极差标准化的基础上，利用夹角余弦法计算相似矩阵，再利用传递闭包法求模糊等价矩阵，对于给定的 $\lambda\in[0,1]$，得到模糊聚类动态聚类图(图 11-1)。若以相似程度 0.75 为标准，18 个洼陷明显地分为两类。根据动态聚类过程，得到第一类洼陷的先后勘探排序(表 11-2)。对于第二类洼陷，根据已有的勘探资料，同样可以得出先后勘探顺序。洼陷的勘探排序可为石油勘探决策提供依据。

表 11-1　胜利油田 18 个洼陷评价指标的专家组评分

洼陷名称	指标名称												
	沉积类型	构造发育	有机质丰度	主生油层埋深	生油岩面积	生油层累计厚度	油岩与泥岩体积比	单位面积生烃量	单位体积生烃量	单位面积排烃量	生储配置关系	运移方式	勘探程度
利　津	5	3	8	10	8	8	4	8	8	6	3	4	8
博　兴	5	3	8	7	8	8	3	4	8	4	3	4	8
牛　六	5	3	8	8	8	8	4	8	8	6	3	4	8
民　风	5	3	8	8	2	8	4	8	8	6	3	3	8
渤　南	4	3	8	10	8	8	4	8	8	6	3	4	8
五号桩	4	2	4	8	2	4	4	8	8	6	3	3	8
孤　南	4	3	8	7	7	4	3	4	4	4	7	4	8
富　林	4	2	4	6	2	4	3	2	4	2	3	3	4
流　钟	2	1	2	3	4	2	1	2	2	2	3	2	2
郭局子	4	2	4	7	2	8	3	8	8	6	3	3	4
大王北	4	2	4	7	7	4	3	8	8	6	3	3	4
车　西	4	2	4	7	8	8	3	4	4	4	3	3	4
临　南	5	3	8	8	8	8	4	4	4	6	3	8	8
磁　镇	2	1	4	3	8	4	1	2	2	2	3	2	2
阳　信	2	1	4	3	8	8	1	2	2	2	3	2	4
里则镇	2	1	2	3	8	2	1	2	2	2	3	2	2
扬　盘	2	1	2	2	4	2	1	2	2	2	3	2	2
滩　北	3	2	4	8	4	8	4	4	2	6	3	3	4

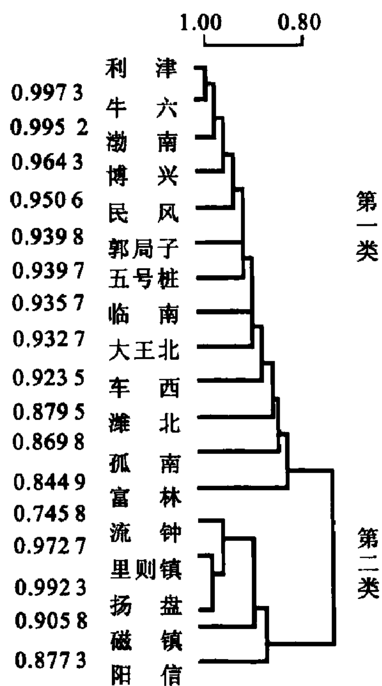


图 11-1　洼陷模糊动态聚类图

表 11-2 第一类有利勘探洼陷排序表

洼陷名称	利 津	牛 六	渤 南	博 兴	民 风	郭局子	五号桩	临 南	大王北	车 西	潍 北	孤 南	富 林
排 序	1	2	3	4	5	6	7	8	9	10	11	12	13

【例 2】气测井资料识别储层含油气性。

设论域U=｛储层含油气性｝上有油层、油水同层、含油水层、干层等模糊模型，待识别含油气性的储层为模糊模型$\boldsymbol{Y}$，试根据贴近度判定$\boldsymbol{Y}$的含油气性。

(1) 建立标准模糊模型库。

在试油证实的油层、油水同层、含油水层、干层内各取若干个样品，每个样品都有相同的气测指标，它们都是模糊变量。把各类样品气测指标的平均值作为标准模糊模型，由它们构成标准模糊模型库(表 11-3)。

表 11-3 储层含油气性标准模型数据

模型库的标准模型	指标平均值						
	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_6$	$\bar{x}_7$
油层($\boldsymbol{X}_1$)	0.011 3	0.012 9	0.014 9	0.200 9	3.147 5	0.052 0	0.105 4
油水同层($\boldsymbol{X}_2$)	0.050 3	0.231 4	0.119 4	0.765 1	1.027 9	0.436 1	2.791 4
含油水层($\boldsymbol{X}_3$)	0.095 9	0.105 0	0.265 0	0.591 3	1.028 3	0.171 6	0.001 9
干层($\boldsymbol{X}_4$)	0.005 4	0.027 1	0.006 8	0.101 8	4.243 0	0.132 2	0.047 6

表 11-3 中，$\bar{x}_i$ 为气测指标$x_i(i=1,2,\cdots,7)$的平均值，它是用逐步判别分析方法从数项气测综合指标中筛选出的识别能力强的综合指标，其定义为：

$$x_1 = c_3/s_1$$

$$x_2 = c_5/s_1$$

$$x_3 = ic_4/c_1$$

$$x_4 = ic_4/c_3$$

$$x_5 = c_1/s_2$$

$$x_6 = c_5/s_2$$

$$x_7 = [58(ic_4 + nc_4) + 72c_5]/(44c_1)$$

$$s_1 = c_1 + c_2 + c_3 + ic_4 + nc_5$$

$$s_2 = c_2 + c_3 + ic_4 + nc_4 + c_5$$

(2) 对待识别储层含油气性进行识别。

计算待识别储层$\boldsymbol{Y}$与$\boldsymbol{X}_i(i=1,2,3,4)$的贴近度$\sigma_0(\boldsymbol{X}_i,\boldsymbol{Y})(i=1,2,3,4)$，若$\sigma_0(\boldsymbol{X}_\alpha,\boldsymbol{Y})$是其中的最大者($\alpha\in(1,2,3,4)$)，则认为待识别储层$\boldsymbol{Y}$的含油性与$\boldsymbol{X}_\alpha$相同。部分储层样品识别效果见表 11-4。

表 11-4　某地区储层含油气性模式识别与试油结果

井　名	深度/m	识别结果	试油结果
Zh101c	3 427	油　层	油　层
Zh101c	3 432	油　层	油　层
潜　山	4 069	油　层	油　层
潜　山	3 984	油水同层	油水同层
潜　山	3 952	油水同层	油水同层
Zh104	3 217	干　层	含油水层
Zh10	4 650	含油水层	含油水层
Zh10	4 660	含油水层	含油水层
Ch307	3 810	干　层	干　层
Zh104	3 887	干　层	干　层
潜　山	3 818	干　层	油水同层

思考与练习

1. 什么是模糊数学？它的研究对象是什么？
2. 什么是模糊相似矩阵和模糊等价矩阵？
3. 什么是模糊矩阵的 λ 截矩阵？如何利用它进行动态聚类？
4. 什么是隶属度和贴近度？它们的含义是什么？
5. 何谓模糊模式识别？用其可以研究地学中的哪些问题？
6. 熟悉模糊相似矩阵和隶属度、贴近度的计算方法。
7. 茶叶等级的识别。

设论域 $U=\{$茶叶$\}$，茶叶等级标准模型为 $\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\boldsymbol{X}_4,\boldsymbol{X}_5$，待识别的茶叶样品为 $\boldsymbol{Y}$，衡量茶叶质量的指标为条索、色泽、净度、汤色、香气和滋味。茶叶等级标准模型与样品的有关数据见表 11-5。

表 11-5　茶叶标准模型与样品数据

质量指标	茶叶标准模型					茶叶样品
	X_1	X_2	X_3	X_4	X_5	
条　索	0.5	0.3	0.2	0.0	0.0	0.4
色　泽	0.4	0.2	0.2	0.1	0.1	0.2
净　度	0.3	0.2	0.2	0.2	0.1	0.1
汤　色	0.6	0.1	0.1	0.1	0.1	0.4
香　气	0.5	0.2	0.1	0.1	0.1	0.5
滋　味	0.4	0.2	0.2	0.1	0.1	0.6

试按照

$$\sigma_0(\boldsymbol{X}_i,\boldsymbol{Y}) = \frac{1}{2}[\boldsymbol{X}_i\circ\boldsymbol{Y} + (1-\boldsymbol{X}_i * \boldsymbol{Y})] \qquad (i=1,2,\cdots,5)$$

计算茶叶样品 $\boldsymbol{Y}$ 与模型 $\boldsymbol{X}_i(i=1,2,\cdots,5)$ 的贴近度，并按择近原则确定茶叶样品 $\boldsymbol{Y}$ 的等级。

第十二章 克立金法简介

§1 随机场与区域化变量

一、随机函数和随机场

若只能知道某变量的取值范围，而不能预测它取何值，则称这种变量为随机变量，如地层水的矿化度、储层的孔隙度和渗透率等。在实际中经常碰到观测结果是依赖于时间参数 t 的函数，并且在相同条件下重复观测时，函数的形式也不一样，如在相同震源下，地震记录的振幅是时间参数 t 的函数，记为 $A(t)$。显然，$A(t)$随 t 而异，$A(t_0)$是一随机变量。又如海浪起伏的振幅可视为依赖于点(x,y)和时间 t 的随机变量。在此基础上给出随机函数和随机场的概念。

1. 随机函数

设 $\Omega=\{\omega\}$是随机实验 E 的样本空间，若对每个 $\omega\in\Omega$ 都有一个函数 $Z(x_1,x_2,\cdots,x_n;\omega)$与之对应($x_i\in X_i,i=1,2,\cdots,n$)，当 x_i 取任意定值 x_{i0}时，$Z(x_{10},x_{20},\cdots,x_{n0};\omega)$是一个随机变量，则称 $Z(x_1,x_2,\cdots,x_n;\omega)$是定义在$(x_1,x_2,\cdots,x_n)$上的一个随机函数。当 $n=1$ 时，称上述随机函数为随机过程，常简记为 $Z(x)$。

2. 随机场

随机场是依赖于多个自变量的随机函数，常用的是以空间点的坐标为自变量的随机场，记为 $Z(x_u,y_v,z_w;\omega)$。在此，以 $\boldsymbol{x}$ 代表三维向量(x_u,y_v,z_w)，并把随机场简记为 $Z(\boldsymbol{x})$。如磁场、重力场等都是空间点的函数，并有不同程度的随机性，故可将其视为随机场。对它们的每次观测结果，都是一个确定的空间点函数，当 $\boldsymbol{x}=\boldsymbol{x}_0$ 时，$Z(\boldsymbol{x}_0)$是个随机变量。

二、区域化变量

1. 区域化变量

区域化变量是以三个直角坐标为自变量的随机场 $Z(\boldsymbol{x})$。对其观测前它是一个随机场，观测后得到它的一个实现，记为 $z(\boldsymbol{x})$。每个 $z(\boldsymbol{x})$是一个三元实函数或空间点函数。

在地质、采矿和油气勘探开发领域，有很多变量可以视为区域化变量，如煤层厚度、生油岩厚度、储集层厚度、有机碳含量。对于这一类的地质变量，可视为区域化变量在二维空间的分布。各种矿化现象是某种元素含量在三维空间的变化。

区域化变量同时反映变量的随机性和结构性，它是地质统计学的研究对象。若 $Z(\boldsymbol{x})$表示储层厚度，那么它是个随机变量，这体现了储层厚度的随机性；另外，在点 $\boldsymbol{x}$ 和 $\boldsymbol{x}+\boldsymbol{h}$($\boldsymbol{h}$ 是二维空间的距离向量)处的储层厚度 $Z(\boldsymbol{x})$和 $Z(\boldsymbol{x}+\boldsymbol{h})$具有一定程度的自相关性。一般来说，$\boldsymbol{h}$ 越小，相关程度越高。自相关性反映了变量的某种连续性和结构性。马特隆在研究金属矿的品位时指出："一个矿床中的矿石品位的分布具有混杂的特征，其中一部分是结构性的，而另一部分则是随机性的……因此，对任何一个矿床进行科学的(至少是符合实际的)估计时，必须既要考虑到矿床固有的结构性，又要考虑到矿床固有的随机性。"

区域化变量反映地质变量的如下特性：

① 局部性。区域化变量只局限于一定的空间内，称该空间为区域化的几何域。

② 连续性。区域化变量具有各自的连续性，也有些变量具有平均的连续性。

③ 异向性。区域化变量在各个方向上的性质不同。

④ 可迁性。它是指区域化变量相关性的局部性。区域化变量在其几何域内具有明显的相关性，但超出它的几何域，相关性明显降低，甚至消失。

上述特性是导致传统资源量估算方法存在较大估算偏差的主要因素，而地质统计学的变差函数，则能较好地研究区域化变量的上述特性。

2. 区域化变量的数字特征

对于油气勘探开发、资源评价等领域的区域化变量，一般不需要知道它的全部特征，只要知道它的某些数字特征就足够了。值得注意的是：与随机变量的数字特征不同，区域化变量的数字特征一般都是函数。以下介绍区域化变量的平均值、方差和协方差函数。

(1) 区域化变量的平均值函数。

区域化变量 $Z(\boldsymbol{x})$的平均值是函数 $E[Z(\boldsymbol{x})]$，对于自变量 $\boldsymbol{x}$ 的每一个给定值 $\boldsymbol{x}_0$，它的函数值等于区域化变量 $Z(\boldsymbol{x})$在 $\boldsymbol{x}_0$ 处的平均值，即

$$E[Z(\boldsymbol{x})]_{x=x_0} = E[Z(\boldsymbol{x}_0)] \tag{12-1}$$

显然，$E[Z(\boldsymbol{x})]$是区域化变量 $Z(\boldsymbol{x})$的所有实现 $z(\boldsymbol{x})$的一个平均函数。

(2) 区域化变量的方差函数。

区域化变量 $Z(\boldsymbol{x})$的方差是函数 $D^2[Z(\boldsymbol{x})]$，对于自变量 $\boldsymbol{x}$ 的每一个给定值 $\boldsymbol{x}_0$，它的函数值等于区域化变量 $Z(\boldsymbol{x})$在 $\boldsymbol{x}_0$ 处的方差，即

$$D^2[Z(\boldsymbol{x})]_{x=x_0} = D^2[Z(\boldsymbol{x}_0)] \tag{12-2}$$

方差函数也可以记为 $\mathrm{Var}[Z(\boldsymbol{x})]$。$D^2[Z(\boldsymbol{x})]$既表示方差，又表示依赖空间点的位置 $\boldsymbol{x}$。

(3) 区域化变量的协方差函数。

协方差函数是随机过程 $Z(t)$中的数字特征。它是 $Z(t)$在时刻 t_1, t_2 处的两个随机变量 $Z(t_1)$和 $Z(t_2)$的二阶混合中心矩，其定义为：

$$\mathrm{Cov}[Z(t_1), Z(t_2)] = E[Z(t_1)Z(t_2)] - E[Z(t_1)]E[Z(t_2)] \tag{12-3}$$

常简记为 $C(t_1, t_2)$，它是随机过程 $Z(t)$的自协方差函数，简称协方差函数。

对于区域化变量 $Z(\boldsymbol{x})$，把上式中 t_1, t_2 变为空间中的两点 $\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}$ 处的两个随机变量 $Z(\boldsymbol{x})$和 $Z(\boldsymbol{x}+\boldsymbol{h})$，仿照式(12-3)可得相应的二阶混合中心矩：

$$\begin{aligned}\mathrm{Cov}[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})] &= C(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) \\ &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] - E[Z(\boldsymbol{x})]E[Z(\boldsymbol{x}+\boldsymbol{h})]\end{aligned} \tag{12-4}$$

式(12-4)称为区域化变量 $Z(\boldsymbol{x})$的自协方差函数，简称协方差函数。

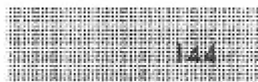
通常 $C(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h})$是依赖于空间点 $\boldsymbol{x}$ 和增量 $\boldsymbol{h}$ 的函数。当 $\boldsymbol{h}=\boldsymbol{0}$ 时，它等于方差函数，即

$$C(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{0}) = E[Z(\boldsymbol{x})]^2 - \{E[Z(\boldsymbol{x})]\}^2 = D^2[Z(\boldsymbol{x})] = \mathrm{Var}[Z(\boldsymbol{x})] \tag{12-5}$$

§2 区域化变量的变差函数及理论模型

一、区域化变量的变差函数

设空间点 $\boldsymbol{x}$ 只在一维 x 轴上变化，定义区域化变量 $Z(x)$在 x 方向上的变差函数为：



$$\begin{aligned}\beta(x,h) &= \text{Var}[Z(x)-Z(x+h)]/2 \\ &= E[Z(x)-Z(x+h)]^2/2-\{E[Z(x)-Z(x+h)]\}^2/2\end{aligned} \quad (12\text{-}6)$$

由上式可知，$\beta(x,h)$是在 $x,x+h$ 两处取值之差的方差的 1/2。若 $\beta(x,h)$只依赖于 h(滞后，距离、间隔、步长)，那么变差函数可记为 $\beta(h)$。以 $\beta(h)$和 h 为纵、横坐标绘出的图形叫做变差图(图 12-1)。变差函数有 a,C,C_0 三个主要参数，其中 a 为变程，反映变量的影响范围；C_0 为块金常数，反映变量的连续性；C_0+C 为基台值；C 为拱高。

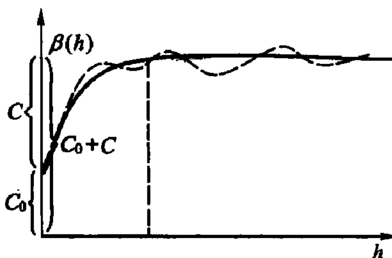


图 12-1 变差函数示意图

变差函数既能描述区域化变量的空间结构性，又能描述其随机性，它是地质统计学所特有的基本工具，也是进行地质统计学计算的基础。当 x 和 h 在二或三维空间变化时，还要考虑二或三维变差函数。

二、平稳假设与本征假设

由式(12-6)可知，欲知 $\beta(x,h)$的估计值，需知 $E[Z(x)-Z(x+h)]^2$ 和 $E[Z(x)-Z(x+h)]$。若用观测值的平均值作为它们的估计值，就必须对 $Z(x)$和 $Z(x+h)$进行若干次观测。但实际工作中不能重复观测，只能得到一对数据 $z(x)$和 $z(x+h)$，因此，无法求出 $\beta(x,h)$的数学期望。为此，对 $Z(x)$给出以下假设。

1. 二阶平稳假设

如果 $Z(x)$满足条件：

① $Z(x)$在整个研究区域内的数学期望存在且为常数，即

$$E[Z(x)] = m, \forall x$$

② $Z(x)$在整个研究区域内的协方差函数存在且相同，即

$$\text{Cov}[Z(x),Z(x+h)] = E[Z(x)Z(x+h)]-m^2 \triangleq C(h), \forall x$$

当 $h=0$ 时，上式变为

$$\text{Cov}[Z(x),Z(x+0)] = \text{Cov}[Z(x),Z(x)] = \text{Var}[Z(x)], \forall x \quad (12\text{-}7)$$

则称 $Z(x)$满足二阶平稳(或弱平稳)假设。

上述结果表明，$Z(x)$在整个研究区域内的协方差函数只依赖于滞后 h，而与 x 无关。当 $h=0$ 时，$Z(x)$在整个研究区域内的方差存在且为常数。

2. 本征假设

若 $Z(x)$在整个研究区域内不满足二阶平稳假设，则提出以下本征假设。

如果 $Z(x)$的增量 $\Delta Z(x)=Z(x)-Z(x+h)$满足条件：

① $\Delta Z(x)$在整个研究区域上的数学期望等于 0，即

$$E[\Delta Z(x)] = 0, \forall x, \forall h$$

② $\Delta Z(x)$的方差函数在整个研究区域上存在且平稳，即

$$\text{Var}[\Delta Z(x)] = E[Z(x)-Z(x+h)]^2, \forall x, \forall h \quad (12\text{-}8)$$

则称 $Z(x)$满足本征假设，或者说 $Z(x)$是本征的。

在平稳假设和本征假设下，增量 $\Delta Z(x)$的数学期望为 0，并与具体位置 x 无关，因此变差函数式(12-6)变为：

$$\beta(h) = E[Z(x)-Z(x+h)]^2/2 \quad (12\text{-}9)$$

上式表明，$Z(x)$在点 x 和 $x+h$ 处的值有差异，$\beta(h)$反映了它们的变异程度。

由式(12-8)，(12-9)得：

$$2\beta(h) = \mathrm{Var}[Z(x) - Z(x+h)] \tag{12-10}$$

本征假设条件②与 $Z(x)$的变差函数存在且平稳是等价的。

从二阶平稳假设与本征假设的定义可知，两者的研究对象不同，前者讨论 $Z(x)$本身的特征，而后者讨论 $Z(x)$增量的特征，并且前者的要求强于后者。

三、实验变差函数的计算公式

它是根据 $Z(x)$的观测值构造的变差函数，记为 $\beta^*(h)$。在二阶平稳假设或本征假设下，可视$[z(x_i), z(x_i+h)]$为$[Z(x_i), Z(x_i+h)]$的不同实现。此时，实验变差函数的计算式为：

$$\beta^*(h) = \frac{1}{2n}\sum_{i=1}^{n}[z(x_i) - z(x_i+h)]^2 \tag{12-11}$$

式(12-11)是观测值规则分布情况下实验变差函数的计算式，其中 n 是被 h 相隔的数据对个数。

把 h_i 代入式(12-11)求出相应的 $\beta^*(h_i)$，并以 h_i，$\beta^*(h_i)$为绘图数据，在 h-$\beta^*(h)$坐标系中绘制一幅折线图(实验变差函数图)。

【例 1】设有甲、乙两个地质单元，在相同方向上测得它们的储层厚度数据(表 12-1)。

表 12-1　地质单元储层厚度数据表

观测点序号	1	2	3	4	5
观测点坐标	500	1 000	1 500	2 000	2 500
甲地质单元储层厚度	2	4	3	1	5
乙地质单元储层厚度	1	2	3	4	5

解：由实测数据可知，甲、乙地质单元储层厚度的均值、方差相同，不能用它们区别甲、乙地质单元储层厚度变化的差异。实际上储层厚度沿观测方向的变化有明显的差异(图 12-2)。

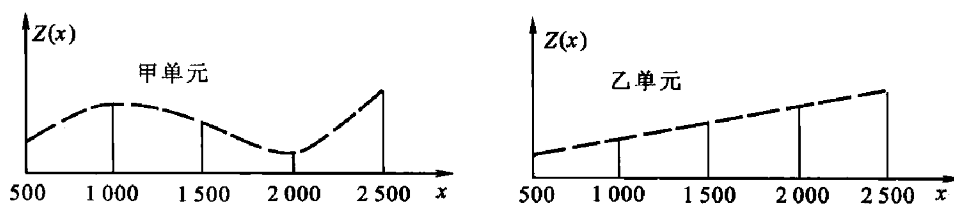


图 12-2　储层厚度沿观测方向变化示意图

采用实验变差函数分析厚度沿观测方向的变化规律。取 $h=500$，据式(12-11)计算实验变差函数值。

甲单元实验变差函数值为：

$$\beta^*(h) = \frac{1}{2\times 4}\sum_{i=1}^{4}[z(x_i) - z(x_i+h)]^2 = 3.1$$

$$\beta^*(2h) = \frac{1}{2\times 3}\sum_{i=1}^{3}[z(x_i) - z(x_i+2h)]^2 = 2.3$$

$$\beta^*(3h)=\frac{1}{2\times 2}\sum_{i=1}^{2}[z(x_i)-z(x_i+3h)]^2=0.5$$

$$\beta^*(4h)=\frac{1}{2\times 1}\sum_{i=1}^{1}[z(x_i)-z(x_i+4h)]^2=4.5$$

同理，乙单元实验变差函数值为：$\beta^*(h)=0.5$，$\beta^*(2h)=2$，$\beta^*(3h)=4.5$，$\beta^*(4h)=8$。

储层厚度实验变差函数图(图 12-3)反映了甲、乙单元储层厚度沿观测方向的变化有明显的差异。甲单元实验变差函数曲线的变化趋势基本上代表了储层厚度沿观测方向的波状起伏。乙单元实验变差函数曲线的变化趋势反映了储层沿观测方向急剧加厚。$\beta^*(h)$与 h 具有良好的相关性，显示了储层厚度的结构性。

【例 2】设生油岩厚度是满足本征假设的二维区域化变量 $Z(x)$，并在矩形区域内得到了它的若干观测值 $z(x)$(图 12-4)。设小正方形的边长为 a，试在 a_1，a_2，a_3，a_4 方向上各计算三个滞后距$|h|$的实验变差函数值，并做出 a_1 和 a_2 方向的实验变差函数图。

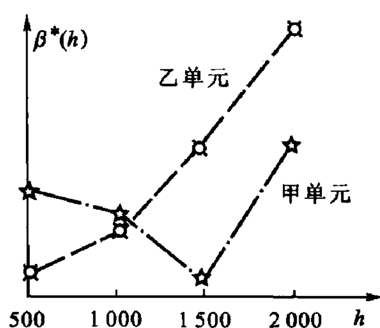


图 12-3 储层厚度实验变差函数图

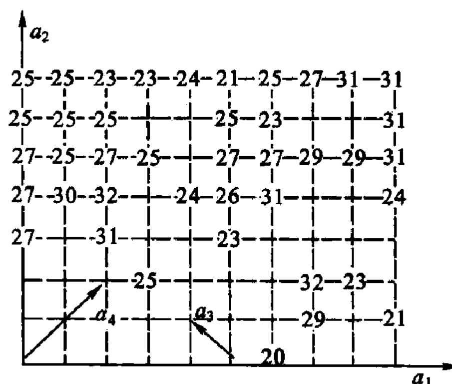


图 12-4 生油岩厚度分布图

解：在每个方向上取三个不同的$|h|$值，根据式(12-11)计算各方向上的 $\beta^*(h)$。

① 方向 a_1($|h|=a$)：

$$\begin{aligned}\beta^*(a)=&[(32-23)^2+(27-30)^2+(30-32)^2+(24-26)^2+(26-31)^2\\&+(27-25)^2+(25-27)^2+(27-25)^2+(27-27)^2+(27-29)^2\\&+(29-29)^2+(29-31)^2+(25-25)^2+(25-25)^2+(25-23)^2\\&+(25-25)^2+(25-23)^2+(23-23)^2+(23-24)^2+(24-21)^2\\&+(21-25)^2+(25-27)^2+(27-31)^2+(31-31)^2]/(2\times 24)\\=&4.10\end{aligned}$$

同理：

$$\beta^*(2a)=336/(2\times 19)=8.84$$

$$\beta^*(3a)=435/(2\times 18)=12.08$$

② 方向 a_2($|h|=a$)：

$$\beta^*(a)=187/(2\times 22)=4.26$$

$$\beta^*(2a)=296/(2\times 18)=8.22$$

$$\beta^*(3a)=327/(2\times 15)=10.90$$

③ 方向 a_3（$|h|=\sqrt{2}a$）：

$$\beta^*(\sqrt{2}a) = 191/(2\times 19) = 5.03$$
$$\beta^*(2\sqrt{2}a) = 381/(2\times 16) = 11.91$$
$$\beta^*(3\sqrt{2}a) = 345/(2\times 10) = 17.25$$

④ 方向 a_4（$|h|=\sqrt{2}a$）：

$$\beta^*(\sqrt{2}a) = 233/(2\times 18) = 6.47$$
$$\beta^*(2\sqrt{2}a) = 315/(2\times 14) = 11.25$$
$$\beta^*(3\sqrt{2}a) = 247/(2\times 8) = 15.44$$

根据上述数据绘出 a_1，a_3 方向的实验变差函数图（图12-5）。

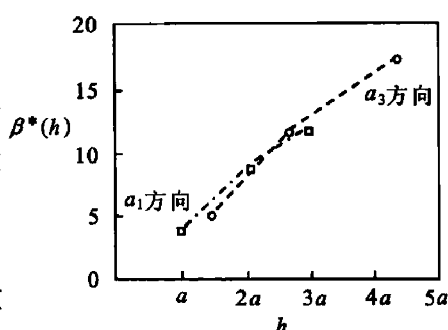


图 12-5　生油岩厚度实验变差函数图

由本例看出，在观测值规则分布的情况下，随着点对之间距离的增大，参与计算的数据对数逐渐减少，这就意味着变差函数曲线上的点距原点越远，它的可靠性越差。

四、变差函数理论模型

为了对区域化变量进行估计，还需将实验变差函数拟合成相应的理论变差函数模型，这些模型将直接参与克立金法的资源量估算以及其他估值。下面讨论一维区域化变量变差函数的理论模型，它们可分为两大类。

1. 有基台值模型

(1) 球状模型。

该模型的一般表达式为：

$$\beta(r) = \begin{cases} 0 & r = 0 \\ C_0 + C\left(\dfrac{3r}{2a} - \dfrac{r^3}{2a^3}\right) & 0 < r \leqslant a \\ C_0 + C & r > a \end{cases} \tag{12-12}$$

式中　C_0——块金常数；

C_0+C——基台值；

C——拱高；

a——变程。

当 $C_0=0$，$C=1$ 时，模型称为标准球状模型（图 12-6）。

(2) 指数函数模型。

该模型的一般表达式为：

$$\beta(r) = C_0 + C(1 - \mathrm{e}^{-r/a}) \tag{12-13}$$

式中　C_0——块金常数；

C_0+C——基台值；

C——拱高；

$3a$——变程。

当 $C_0=0$，$C=1$ 时，模型称为标准指数模型（图 12-6）。

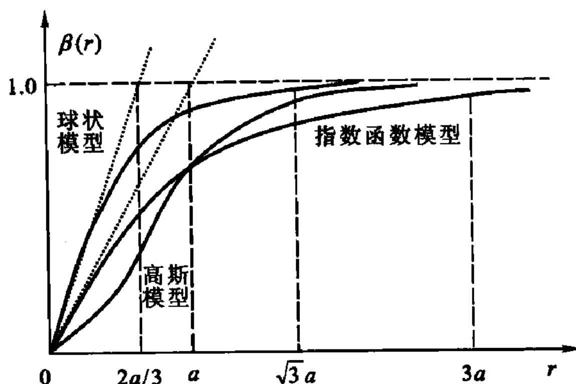


图 12-6 变差函数模型

(3) 高斯模型。

该模型的一般表达式为:

$$\beta(r) = C_0 + C(1 - e^{-r^2/a^2}) \tag{12-14}$$

式中 C_0——块金常数;

C_0+C——基台值;

C——拱高;

$\sqrt{3}a$——变程。

当 $C_0=0, C=1$ 时,称模型为标准高斯模型(图 12-6)。

2. 无基台值模型

(1) 幂函数模型。

该模型的一般表达式为:

$$\beta(r) = r^\theta \quad 0 < \theta < 2 \tag{12-15}$$

常用的是 $\theta=1$ 的线性模型(图 12-7),即

$$\beta(r) = \omega r \tag{12-16}$$

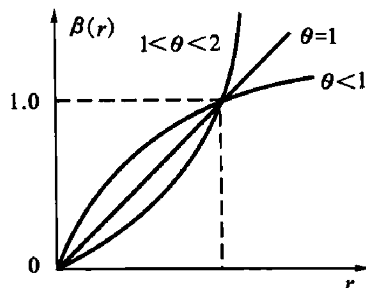


图 12-7 幂函数模型

式中 ω 为一常数,表示直线的斜率。对于很小的 $|h|$,可用线性拟合在原点处具有线性状的任一模型(如球状模型)。

(2) 纯块金效应模型。

该模型的一般表达式为:

$$\beta(r) = \begin{cases} 0 & r = 0 \\ C_0 & r > 0 \end{cases} \tag{12-17}$$

该模型适合于对纯随机变量的估计。可视其 a 为无穷小量,$C=0$,对任何 $r>0$,$\beta(r)$ 都能达到基台值 C_0。

§3 实验变差函数的拟合与结构叠合

一、实验变差函数的拟合

无论是要了解 $Z(x)$ 的变异特征,还是要对其进行估算,都必须知道变差函数的理论模型以及模型中的参数。为此,首先要根据 $Z(x)$ 的观测值 $z(x_i)(i=1,2,\cdots,n)$,按式(12-11)求出实验变差函数,再根据实验变差函数选择适当的理论变差函数进行拟合,从而求出各个

参数。实际工作中的大部分变差函数模型是球状模型，在此假设下，讨论实验变差函数的拟合。

1. 直接拟合

（1）求变程 a。计算观测值的实验方差

$$\sigma^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}[z(x_i)-\bar{z}]^2$$

式中 $\bar{z}=\frac{1}{n}\sum_{i=1}^{n}z(x_i)$。

在实验变差函数图（图 12-8）的纵坐标轴上过 σ^{*2} 点作平行于横轴的直线。设直线 l 与 $\beta^*(h)$ 的头部有 2～3 个交点，并与过 σ^{*2} 点的直线相交，交点的横坐标为 $2a/3$，假定其值为 h_0，则 $a=3h_0/2$。

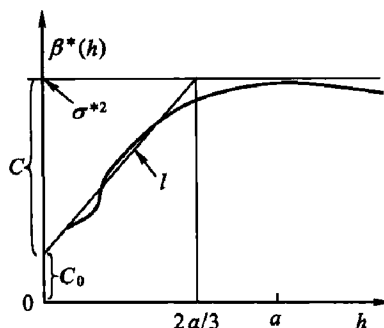


图 12-8　求参数示意图

（2）求块金值。直线 l 与纵轴交点的纵坐标为块金值 C_0。当 $C_0<C$ 时，取 $C_0=0$。

（3）求拱高值。拱高 $C=\sigma^{*2}-C_0$。

2. 多项式回归拟合球状模型

对于球状模型 $0<h\leqslant a$ 的情况，令

$$y=\beta(h),\ x_1=h,\ x_2=h^3,\ b_0=C_0,\ b_1=\frac{3C}{2a},\ b_2=-\frac{C}{2a^3}$$

则有 $y=b_0+b_1x_1+b_2x_2$，利用线性回归方法可求得上式中的 b_0,b_1,b_2，进而求出球状模型。

3. 交叉验证

交叉验证即检验确定的变差函数是否符合实际，用统计的术语就是估计值与真实值的误差平方和是否最小。具体方法是：对于每个实测点，用其周围点上的值对该点进行克立金估值。若有 n 个实测点，则各有 n 个实测值和克立金估计值，其误差平方为 $E=(Z^*-Z)^2$，误差平方的均值为 $\bar{E}$。$\bar{E}$ 越小，拟合的变差函数越好。

二、实验变差函数的结构叠合

$Z(x)$ 的变异性通常包含各种尺度上的多层次的变异性，反映在变差函数上，就是其结构性往往不是单一的，而是多层次结构的叠加，这种变化结构称为叠合结构。大尺度的变异总是包含小尺度的变异，但是却不能从大尺度的变异性中区分出小尺度的变异性。实验变差函数结构叠合的目的是获得一个随着距离和方向变化的变差函数。

1. 各向同性条件下的结构叠合

各向同性是指 $Z(x)$ 在不同方向上的变异性相同。此时，可直接将变差函数叠合，即

$$\beta(r)=\beta_0(r)+\beta_1(r)+\cdots$$

不同尺度上的结构可以是相同模型，也可以是不同模型。若某各向同性 $Z(x)$ 的变差函数由

$$\beta_0(r)=\begin{cases}0 & r=0\\ C_0 & r>0\end{cases}$$

$$\beta_1(r)=\begin{cases}C_1\left(\frac{3r}{2a_1}-\frac{r^3}{2a_1^3}\right) & 0\leqslant r\leqslant a_1\\ C_1 & r>a_1\ (a_1<a_2)\end{cases}$$

$$\beta_2(r)=\begin{cases}C_2(\dfrac{3r}{2a_2}-\dfrac{r^3}{2a_2^3}) & 0\leqslant r\leqslant a_2\\ C_2 & r>a_2\end{cases}$$

三个不同尺度上的结构组成，则 $Z(x)$ 的变差函数为上述变差函数的和，即

$$\beta(r)=\begin{cases}0 & r=0\\ C_0+3/[2(C_1/a_1+C_2/a_2)r]-1/[2(C_1/a_1^3+C_2/a_2^3)r^3] & 0<r\leqslant a_1\\ C_0+C_1+C_2(\dfrac{3r}{2a_2}-\dfrac{r^3}{2a_2^3}) & a_1<r\leqslant a_2\\ C_0+C_1+C_2 & r>a_2\end{cases}\tag{12-18}$$

叠合过程如图 12-9。

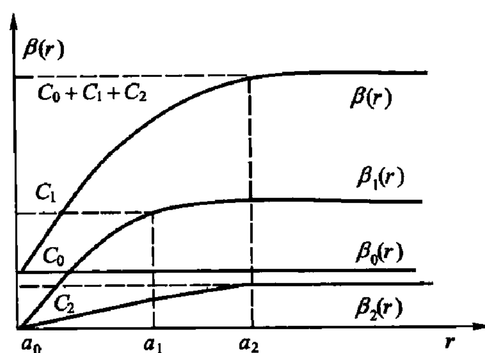
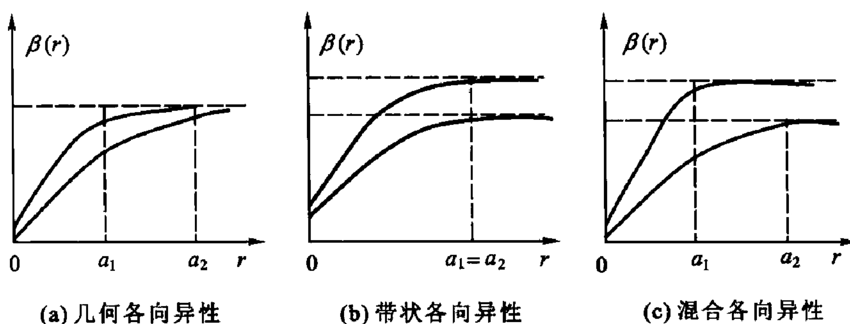


图 12-9　各向同性叠合过程示意图

2. 各向异性条件下的结构叠合

各向异性是指 $Z(x)$ 在不同方向上的变异性不同，它又分为几何、带状和混合各向异性。几何各向异性的基台值在各方向上相同，但变程不同(图 12-10a)；带状各向异性与几何各向异性相反(图 12-10b)；混合各向异性的基台值和变程在各方向上都不同(图 12-10c)。



(a) 几何各向异性　　(b) 带状各向异性　　(c) 混合各向异性

图 12-10　不同类型各向异性变差图

(1) 几何各向异性下的结构叠合。

当 $Z(x)$ 为几何各向异性时，要先对同一尺度上的各向异性结构进行各向同性化处理，然后再把同性化后的不同尺度上的结构进行叠合。

① 变差函数的各向同性化变换。

设区域化变量在 x 和 y 方向上的变差函数分别为：

$$\beta(h_x)=\begin{cases}0 & h_x=0\\ C(\frac{3h_x}{2a_x}-\frac{h_x^3}{2a_x^3}) & 0<h_x\leqslant a_x\\ C & h_x>a_x\end{cases}$$

$$\beta(h_y)=\begin{cases}0 & h_y=0\\ C(\frac{3h_y}{2a_y}-\frac{h_y^3}{2a_y^3}) & 0<h_y\leqslant a_y\\ C & h_y>a_y\end{cases}$$

那么，它们的基台值均为 $C(0)$，在两个方向上的变程分别为 a_x 和 a_y，属于几何各向异性。通过式

$$\boldsymbol{h}=\begin{bmatrix}h_x/a_x\\ h_y/a_y\end{bmatrix},\qquad h=\sqrt{(h_x/a_x)^2+(h_y/a_y)^2}$$

对坐标变换后，得到一个各向同性化的变差函数

$$\beta(h)=\begin{cases}0 & h=0\\ C(3h/2-h^3/2) & 0<h\leqslant 1\\ C & h>1\end{cases}\tag{12-19}$$

注意：只有当 $\beta(h_x)$，$\beta(h_y)$的函数类型相同时，才能进行上述坐标变换。

② 不同尺度上的结构叠合。

在对同一尺度上的几何各向异性同性化后，可对不同尺度上的结构进行叠合，得：

$$\beta(h)=\beta_1(h_1)+\beta_2(h_2)\tag{12-20}$$

式(12-20)是如下

$$\beta_1(h_1)=\beta_1\ \sqrt{(h_x/a_{x1})^2+(h_y/a_{y1})^2}$$

$$\beta_2(h_2)=\beta_2\ \sqrt{(h_x/a_{x2})^2+(h_y/a_{y2})^2}$$

两个各向同性化之后的变差函数的叠加。

(2) 带状各向异性下的结构叠合。

带状各向异性的变差函数可以看做一种不同方向结构的叠合结构。在此结构中，可直接把各部分相加，即

$$\beta(h)=\sum_{i=1}^{n}\beta_i(h)\tag{12-21}$$

(3) 混合各向异性下的结构叠合。

在混合各向异性中，每一个组成结构可以具有各自不相同的异向性，但都可以通过适当的线性变换化为各向同性后叠加，其中有的 $\beta_i(h)$可以是几何异向性的，对每个组成结构分别进行一种适当的坐标线性变换，把它们转化为各向同性结构，最后再把这些各向同性结构叠加，构成一个统一的变差函数模型。例如，对于在水平和垂直方向上变程为 a_x,a_y,a_z 的区域化变量，可以把水平方向各向同性化后，再加上垂直方向的结构(图 12-11)。

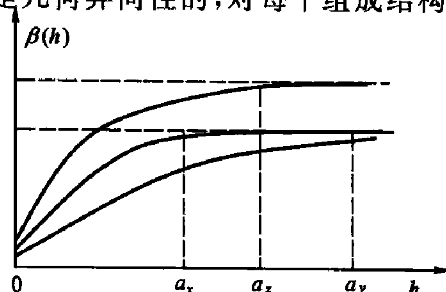


图 12-11　混合各向异性变差函数示意图

水平方向各向同性化后的结构为 $\beta_1(h_1)=\beta_1[(h_x/a_x)^2+(h_y/a_y)^2]^{\frac{1}{2}}$；垂直方向的结构为 $\beta_2(h_2)$

$=\beta_2(h_z/a_z)$；叠合后的结构为 $\beta(h)=\beta_1(h_1)+\beta_2(h_2)$。

§4 克立金法

经过不断发展和完善，逐渐产生了适合不同情况的克立金法。从资源预测的角度来说，多数克立金法是根据一个区域内外若干样品的某种特性的观测值，对该区域同类特性进行线性无偏、最小方差估计的方法。具体地说，该方法是在考虑了样品的形状、大小及其与待估域之间的空间分布位置等几何特征，以及变量的空间结构信息后，为了达到线性无偏和最小估计方差，而对每一个观测值分别赋予一定的权系数，最后用加权平均的方法估计待估域的同类特性。

一、线性估计量

设 $z(v_i)(i=1,2,\cdots,n)$ 是 $Z(x)$ 在信息域 $v_i(i=1,2,\cdots,n)$ 上的观测值，待估域 V 上的真值为 Z_V，x 为 V 的中心。通常来说，Z_V 的估计量 Z^* 是诸 $z(v_i)$ 的函数，将该函数记为：

$$Z^* = f[z(v_1),z(v_2),\cdots,z(v_n)] \tag{12-22}$$

我们期望在无偏性和最优性（估计方差最小）的条件下来确定 f。在此，仅取 f 为线性函数的形式，即

$$Z^* = \sum_{i=1}^{n}\lambda_i z(v_i) = \sum_{i=1}^{n}\frac{\lambda_i}{v_i}\int_{v_i}Z(t)\mathrm{d}t \tag{12-23}$$

在平稳假设条件下采用线性估计量，可以求出 $E(Z_V-Z^*)$ 和 $E(Z_V-Z^*)^2$，由此可以进一步讨论在无偏性和最优性条件下，如何确定式(12-23)中的权系数 λ_i。

二、无偏性条件和线性估计方差

1. 无偏性条件

采用式(12-23)，在满足无偏性和最优性条件下，可以求出权系数 λ_i，得到线性估计量。无偏性条件要求 $E(Z_V-Z^*)=0$，由此分 $E(Z(x))$ 为已知和未知两种情况进行。

(1) $E[Z(x)]$ 为已知的无偏性条件。

设 $E[Z(x)]=m$（常数），令 $Z'(x)=Z(x)-m$，因 $Z(x)$ 为二阶平稳区域化变量，故 $Z'(x)$ 也是二阶平稳区域化变量，且 $E[Z'(x)]=0$。若能对 $Z'(x)$ 在待估域 V 中的变量 Z^* 值进行估计，就可以通过 $Z(x)=Z'(x)+m$ 对 $Z(x)$ 在待估域 V 中的变量 Z^* 值做出估计。为此，把式(12-23)改写为：

$$Z'^* = \sum_{i=1}^{n}\lambda_i z'(v_i) = \sum_{i=1}^{n}\frac{\lambda_i}{v_i}\int_{v_i}Z'(t)\mathrm{d}t \tag{12-24}$$

可以证明，当 $E(Z(x))=m$ 时，满足无偏性条件。

(2) $E[Z(x)]$ 为未知的无偏性条件。

因为 Z_V 和 Z^* 都是随机变量，可以证明，要使 $E(Z_V-Z^*)=0$，必须使

$$\sum_{i=1}^{n}\lambda_i = 1$$

2. 满足无偏性条件时线性估计方差的计算公式

当满足无偏性条件时，可以证明估计方差的计算公式为：

$$\sigma_E^2 = \bar{C}(V,V) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\bar{C}(v_i,v_j) - 2\sum_{i=1}^{n}\lambda_i\bar{C}(V,v_i) \tag{12-25}$$

式中

$$\bar{C}(V,V)=\frac{1}{V^2}\int_V\int_V C(y-y')\mathrm{d}y\mathrm{d}y'$$

$$\bar{C}(V,v_i)=\frac{1}{Vv_i}\int_V\int_{v_i} C(y-t)\mathrm{d}y\mathrm{d}t$$

$$\bar{C}(v_i,v_j)=\frac{1}{v_iv_j}\int_{v_i}\int_{v_j} C(t-t')\mathrm{d}t\mathrm{d}t'$$

其中 $\bar{C}(V,V)$代表距离向量的两个端点各自独立地扫描过整个 V 时，所得变差函数的平均值；$\bar{C}(v_i,v_j)$代表距离向量的两个端点各自独立地扫描过 v_i,v_j 时，所得变差函数的平均值；$\bar{C}(V,v_i)$代表距离向量的两个端点各自独立地扫描过 V,v_j 时，所得变差函数的平均值。

用 v 表示所有的 v_i，则有：

$$\bar{C}(V,v)=\sum_{i=1}^{n}\lambda_i\bar{C}(V,v_i),\quad \bar{C}(v,v)=\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{C}(v,v)$$

则估计方差的计算公式可简写为：

$$\sigma_E^2=\bar{C}(V,V)+\bar{C}(v,v)-2\bar{C}(V,v) \tag{12-26}$$

3. 关于估计方差的说明

(1) 用域 v 内的信息对域 V 作估计求得的 σ_E^2，也称为 v 对 V 的外延方差，记为 $\sigma_E^2(v,V)$。

(2) 不论 v 和 V 是怎样的域，式(12-26)都适用。

(3) 可以把变差函数

$$\beta(h)=E[Z(x)-Z(x+h)]^2/2,\ \forall x$$

视为用 $Z(x+h)$估计 $Z(x)$时的估计方差的一半。

(4) σ_E^2 是衡量估计量优劣的一个指标，σ_E^2 越小，估计量越好。

(5) 影响 σ_E^2 大小的因素有四个方面：

① $\beta(h)$反映了 $Z(x)$的结构特征和空间连续性，因此它的数学模型是影响因素之一。

② V 的几何特征，如大小、形状等。总的来说，V 越大，$\beta(V,V)$也越大，于是 σ_E^2 则相应变小。

③ v 的几何特征、数量和空间排列等。一般说来，v 越大，样品越多，距离越远，$\beta(v,v)$则越大，σ_E^2 则越小。

④待估域与信息样品之间的距离越大，$\beta(V,v)$就越大，因此 σ_E^2 就越大。

三、克立金法

1. 简单克立金法

它是二阶平稳假设下的一种线性无偏估计。当 $E[Z(x)]=m$(已知常数)时，无偏性条件自然满足，由此转化为在估计方差最小条件下求取式(12-24)中的参数 λ_i。对式(12-25)中诸 λ_i 求偏导数，并令其等于 0，化简整理后得：

$$\sum_{j=1}^{n}\lambda_i\bar{C}(v_i,v_j)=\bar{C}(V,v_i)\qquad(i=1,2,\cdots,n) \tag{12-27}$$

由方程组(12-27)可以解出权系数 $\lambda_i(i=1,2,\cdots,n)$，代入式(12-24)对 Z'^* 进行估计，进而由 $Z(x)=Z'(x)+m$ 估计 Z^*。

2. 普通克立金法

它是本征假设下的一种线性无偏最优估计。无偏是要求权系数的和为1,最优是要求达到估计方差最小。这属于满足无偏条件下求条件极值的问题。为此,构造一个新的函数

$$F = \sigma_E^2 - 2\mu(\sum_{i=1}^{n}\lambda_i - 1) \qquad (12\text{-}28)$$

式中 μ——拉格朗日算子。

对式(12-28)中诸 λ_i 和 μ 求偏导数,并令其等于0,化简整理后得普通克立金方程组

$$\begin{cases} \sum_{j=1}^{n}\lambda_i\bar{C}(v_i,v_j) - \mu = \bar{C}(V,v_i) \\ \sum_{i=1}^{n}\lambda_i = 1 \end{cases} \quad (i = 1,2,\cdots,n) \qquad (12\text{-}29)$$

由式(12-29)可以解出权系数 λ_i 和 μ,把 λ_i 代入式(12-23)可求得估计域 V 中变量的估计值 Z^*。

把 λ_i 代入式(12-26)化简整理,得到估计方差的简化表达式,将其称为普通克立金方差,记为:

$$\sigma_{0k}^2 = \bar{C}(V,V) - \sum_{i=1}^{n}\lambda_i\bar{C}(V,v_i) + \mu \qquad (12\text{-}30)$$

3. 指示克立金法

它是针对非参数和无分布的区域化变量提出的一种克立金法。它可以在保留实际存在的高值数据条件下,处理各种不同的现象,并给出一定概率条件下变量的估计值及其空间分布。

(1) 指示函数。

对于任意给定的阈值 z_k,对每个样品点 x_i 定义指示函数

$$I(x_i;z_k) = \begin{cases} 0 & z(x_i) > z_k \\ 1 & z(x_i) \leqslant z_k \end{cases} \quad (k = 1,2,\cdots,K) \qquad (12\text{-}31)$$

指示函数 $I(x_i;z_k)(k=1,2,\cdots,K)$ 的物理含义是样品点 x_i 处随机变量 $Z(x)$ 的概率分布函数,即

$$F(x_i;z_k) = p \mid Z(x_i) \leqslant z_k \mid = I(x_i;z_k) = \begin{cases} 0 & z(x_i) > z_k \\ 1 & z(x_i) \leqslant z_k \end{cases} \quad (k = 1,2,\cdots,K) \qquad (12\text{-}32)$$

(2) 指示协方差函数和指示变差函数。

视指示函数为区域化变量,以 h 代表点 x 和 $x+h$ 之间的距离,仿照区域化变量的协方差函数和变差函数,定义指示协方差函数和指示变差函数为:

$$\begin{aligned} C_I(h;z_k) &= \mathrm{Cov}[I(x;z_k), I(x+h;z_k)] \\ &= E\{I(x;z_k) - E[I(x;z_k)]\}\{I(x+h;z_k) - E[I(x+h;z_k)]\} \\ &\quad (\forall x, \forall h, k = 1,2,\cdots,K) \end{aligned} \qquad (12\text{-}33)$$

$$\beta(h;z_k) = \mathrm{Var}[I(x;z_k) - I(x+h;z_k)]/2 \quad (k = 1,2,\cdots,K) \qquad (12\text{-}34)$$

在二阶平稳假设下,指示协方差函数和指示变差函数的形式为:

$$C_I(h;z_k) = \mathrm{Cov}[I(x;z_k), I(x+h;z_k)] = E[I(x;z_k)I(x+h;z_k) - \mu^2] \quad (k=1,2,\cdots,K) \tag{12-35}$$

$$\beta_I(h;z_k) = E[I(x;z_k) - I(x+h;z_k)]^2/2 \quad (k=1,2,\cdots,K) \tag{12-36}$$

在二阶平稳假设下，指示协方差函数和指示变差函数的关系为：

$$\beta_I(h;z_k) = C_I(0;z_k) - C_I(h;z_k) \quad (k=1,2,\cdots,K) \tag{12-37}$$

实验变差函数的计算式为：

$$\beta_I^*(h;z_k) = \frac{1}{2n}\sum_{i=1}^{n}[I(x_i;z_k) - I(x_i+h;z_k)]^2 \quad (k=1,2,\cdots,K) \tag{12-38}$$

式中 n——相距 h 的数据点对的个数。

指示变差函数拟合的原理和过程与区域化变量变差函数拟合的原理和过程相同。

(3) 指示克立金方程组。

假设变量无空间结构性，那么 x 点处随机变量 $Z(x)$ 的概率分布函数为：

$$F(x;z_k/n) = p[Z(x) \leqslant z_k/n] = \frac{1}{n}\sum_{i=1}^{n} I(x_i;z_k) \quad (k=1,2,\cdots,K) \tag{12-39}$$

式中，F 是 z_k 和 $Z(x_i)(i=1,2,\cdots,n)$ 的函数。

假设变量存在空间结构性，那么估计 x 点处 $Z(x)$ 的概率分布函数则需引用指示克立金法。

当 $Z(x)$ 存在空间结构性时，则有：

$$F(x;z_k/n) = p[Z(x) \leqslant z_k/n] = \sum_{i=1}^{n}\lambda_i(x;z_k)/(x;z_k) \quad (k=1,2,\cdots,K) \tag{12-40}$$

式(12-32)和(12-40)表明，求 $F(x;z_k/n)$ 可以看做用已知的 $I(x;z_k)$ 来估计未知的 $I(x;z_k)$。由概率理论可以证明 $\sum_{i=1}^{n} I(x_i;z_k)$ 确实是已知 n 个数据时 $Z(x)$ 的条件概率分布函数，或是 $I(x;z)$ 的线性估计，即

$$F(x;z_k/n) = I^*(x;z_k) = \sum_{i=1}^{n}\lambda_i(x;z_k)/(x;z_k) \quad (k=1,2,\cdots,K) \tag{12-41}$$

给定任意一点 x 和 z_k，可建立指示克立金方程组

$$\begin{cases} \sum_{j=1}^{n}\lambda_j(x;z_k)\beta_I(x_i - x_j;z_k) + \mu(x;z_k) = \beta_I(x - x_i;z_k) \\ \sum_{j=1}^{n}\lambda_j(x;z_k) = 1 \quad (i=1,2,\cdots,n) \end{cases} \tag{12-42}$$

式中 β_I——指示变差函数；

$x_i - x_j$——两个样品点之间的距离；

$x - x_i$——待估点 x 与样品点 x_i 之间的距离。

由式(12-42)可以解出系数 $\lambda_j(x;z_k)$ 和 $\mu(x;z_k)$，代入式(12-41)便可求出 $I^*(x;z_k)$。

通常把 $z(x_i)(i=1,2,\cdots,n)$ 分成 K 个阈值 $z_1, z_2, \cdots, z_K$，因此在每一点 x 上有 K 个指

示克立金方程组需要求解，从而可得到 K 个 $I^*(x;z_k)$ 估计值，它们分别对应待估点 x 处 $Z(x)$ 小于 K 个阈值 $z_1,z_2,\cdots,z_K$ 的概率。

由上可知，在不同的点上可得到不同的概率分布曲线，从这些概率分布曲线上可以读出 $Z(x)$ 在不同点出现小于某个定值 z_0 的概率，由此可绘出等概率图。也可以根据不同点的概率分布曲线，事先确定一个概率值 p_0，绘制变量等值线图，在该图上可以读出在概率 p_0 下 $Z(x)$ 的值。这两种图形皆可反映区域化变量的随机性。

(4) 几点说明。

① 指示克立金法给出 $Z(x)$ 在空间某点处的概率分布曲线，由此既可对 $Z(x)$ 的不确定性进行度量，又可给出 $Z(x)$ 的值大于(或小于)某个给定值的概率，这一结果可用于资源量预测一类的问题。

② 普通克立金法用克立金估计方差描述不确定性，而指示克立金法则用概率来描述。

③ 指示克立金法是一种非线性、非参数估计方法。

④ 指示克立金法适用于连续型和离散型 $Z(x)$。

§5 应用实例

【例 1】普通克立金法应用。

收集了塔里木盆地轮南凸起中部(图 12-12)48 口井的压力数据，在结构分析的基础上，对其中 15 口井的压力用普通克立金法进行了估计，并与实测值进行了对比-交叉验证(表 12-1，据康永尚)。其中，10 口井的预测精度大于 90%，3 口井的预测精度在 88%～90%之间，2 口井的预测精度在 86%～88%之间，结果表明，克立金法预测精度能满足压力预测的要求。

表 12-1 塔里木盆地轮南凸起中部超压压力系统普通克立金法估计值与实测值

井 号	实测值	预测值	绝对误差	预测精度/%
轮南 39	1.143	1.088	0.055	93.8
轮南 54	1.070	1.070	0.000	100.0
轮南 46	1.105	1.085	0.020	98.2
轮南 32	1.212	1.219	0.070	99.4
轮南 21	1.161	1.132	0.029	97.5
轮南 23	1.264	1.251	0.013	99.0
轮南 19	1.226	1.249	0.023	98.1
轮南 16	1.206	1.063	0.142	88.2
轮南 53	1.206	1.112	0.094	92.2
轮南 51	1.247	1.076	0.171	86.3
轮南 11	1.261	1.201	0.060	95.2
轮南 59	1.355	1.216	0.139	89.7
解放 126	1.206	1.073	0.133	89.0
吉 108	1.353	1.268	0.065	93.7
解放 124	1.275	1.107	0.168	86.8

预测精度稍低的轮南51井和解放124井，位于桑塔木断裂带附近(图 12-12)。在断裂带附近，地层的孔渗性得到改善，而断裂带又常常是泄压的通道，故在断裂带附近压力分布得到了一定程度的调整。本方法在进行压力预测时，所能考虑的是地层压力区域化变量在其空间域中的整体结构性状，而在断裂带附近地层压力受局部干扰所出现的局部性结构特点无法考虑，这就导致本方法在断裂带附近预测的精度要低一些。

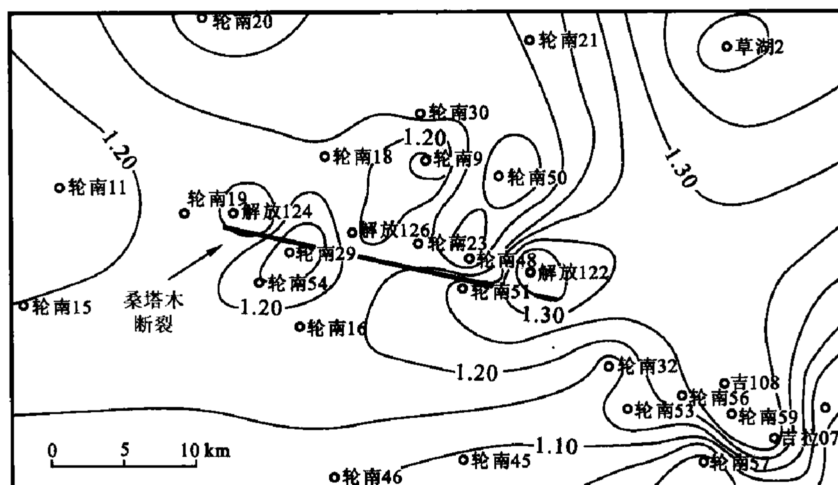


图 12-12 塔里木盆地轮南凸起中部超压压力系统
实测地层压力系统分布和桑塔木断裂位置示意图

【例 2】指示克立金法应用(据黄竞先)。

在处理物化探数据时，常会遇到一些特异值，所谓特异值是指那些比全部数值的均值高得多的数值，这种数值实际存在于所研究的母体之中，虽然它们只占全部数据的极小部分，但在数值估计中却起着很大作用，它们常使数值过高估计。此外，在区域找矿中，人们更感兴趣的与其说是元素的含量，还不如说是大于某一指定异常下限的概率。非参数地质统计学——指示克立金法在解决特异值及区域化变量的概率估计上是有效的。

我们曾经应用指示克立金法处理福建某地区水系沉积物 Mn 元素含量数据，首先根据样品含量的累积频率分布图，分别求出 0.1，0.2，0.3，…，0.9 共 9 个分位数所对应的一组下限值($\times 10^{-6}$)为 300，400，500，600，700，800，1 000，1 500，然后根据各下限值计算半变异函数及进行指示克立金估计，绘制了展示 Mn 含量的指示克立金立体图(图 12-13)和克立金估值概率等值线图(图 12-14)。从概率等值线图上可以清晰地看出区域内大于指定下限值的高概率的地段分布位置。

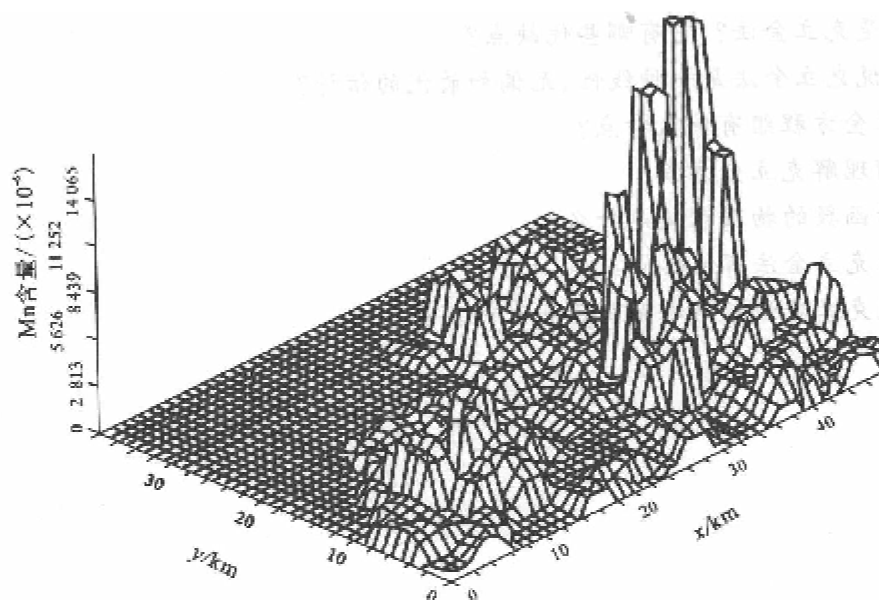


图 12-13 Mn 含量指示克立金估值立体图

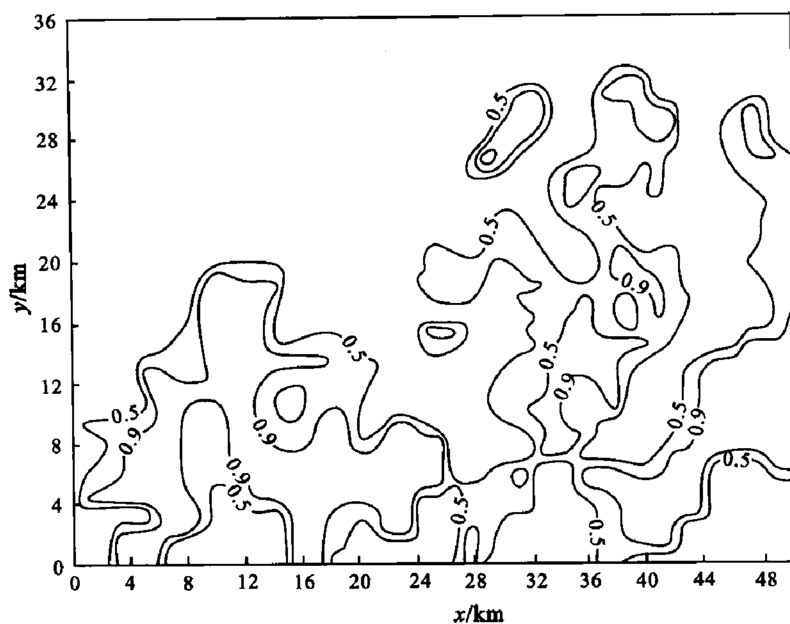


图 12-14 Mn 含量大于 600×10^{-6} 的概率等值线图

思考与练习

1. 如何理解随机函数的概念？
2. 什么是区域化变量和它的二重性？
3. 何谓平稳假设和本征假设？两者有何区别？为什么要做这种假设？
4. 二阶平稳假设条件下的变差函数有何性质和功能？
5. 区域化变量有何地质特征？
6. 如何理解几何各向异性和带状各向异性？
7. 变差函数有哪些基本模型？在各向异性条件下如何进行叠合？

8. 什么是克立金法？它有哪些优缺点？
9. 为何说克立金法是一种线性、无偏和最优的估计？
10. 克立金方程组有什么特点？
11. 如何理解克立金方差？
12. 指示函数的物理意义是什么？
13. 指示克立金法有何优点？
14. 普通克立金法与指示克立金法有何差异？

第十三章 人工神经网络及其应用

§1 人工神经网络

一、人工神经网络概述

长期以来人类一直试图了解和揭示人脑的工作机理和思维的本质,并渴望可以重新构造人脑,这个目标的意义不言而喻。在这个目标的驱使下,20 世纪 40 年代问世的电子计算机实现了人类利用机器模拟人脑逻辑思维的追求。现在计算机的计算速度远远超过了人脑,可以高速求解可编程问题,在数值计算和逻辑运算方面扩展了人脑的能力。然而在解决与形象思维相关的问题时,计算机却显得无能为力。比如欣赏一幅画,人脑可以根据直觉和经验轻松地做到,而对计算机来讲却相当艰巨。

为了模拟人脑的形象思维,人工神经网络应运而生。逻辑思维的模拟可以认为是从心理上对智能进行模拟,是传统的人工智能技术。形象思维的模拟是从生理的角度对智能进行模拟,是基于人工神经网络的人工智能技术。为叙述简便,常将人工神经网络简称为神经网络或网络。

1. 人工神经网络的提出

人工神经网络(artificial neural networks,ANN)是对人类大脑系统的一阶特性的一种描述。简单地讲,它是一个数学模型,可以用电子线路来实现,也可以用计算机程序来模拟,属于人工智能的范畴。首先介绍一些人工智能的概念。

(1) 人工智能。

智能是个体认识客观事物和运用知识解决问题的能力,人工智能就是研究如何让计算机模仿人脑进行工作。

人工智能(artificial intelligence,AI)于 1956 年问世,是一门由计算机科学、控制论、信息论、语言学、神经生理学、心理学、数学、哲学等多种学科相互渗透而发展起来的综合性新学科。

对于人工智能的研究,由于研究角度的不同,形成了不同的研究学派,主要有符号主义学派、连接主义学派和行为主义学派。符号主义又称为逻辑主义、心理学派或计算机学派,其原理主要为物理符号系统(即符号操作系统)假设和有限合理性原理。连接主义又称为仿生学派或生理学派,其主要原理为神经网络及神经网络间的连接机制与学习算法。行为主义又称为进化主义或控制论学派,其原理为控制论及感知-动作型控制系统。

(2) 人工神经网络的提出。

人工神经网络源于人类对人脑的研究。人类对自身思维的好奇产生了许多关于思维的推测,直到神经解剖学家和神经生理学家提出人脑的"通信连接"机制,人类才对人脑有了较深刻的认识。20 世纪 40 年代初期,在对神经元的功能及其功能模式的研究取得一定成果后,研究人员通过这些成果建立起一个数学模型来检验他们的猜想,这标志着人工神经网络的出现。

1943 年生理学家 McCulloch 和数理逻辑学家 Pitts 创立了阈值加权和模型，即 MP 模型，开创了用电子装置模仿人脑结构和功能的新途径。这种模型从研究神经元开始，进而研究神经网络模型和脑模型，开辟了人工智能的又一发展道路，为用元器件和计算机程序实现人工神经网络奠定了基础。

1949 年心理学家 Hebb 发表了《行为构成》一书，提出了现在称为 Hebb 学习律的连接权训练算法。Hebb 也是最早提出“连接主义”的学者之一，这一名词的含义是大脑的活动靠脑细胞的组合连接实现。Hebb 认为如果源神经元和目的神经元均被激活，它们之间突触的连接强度就会增强。这被认为是最早最著名的 Hebb 学习律的生理学基础。Hebb 学习律在人工神经网络的发展史上具有重要地位，是人工神经网络学习训练算法的里程碑。

2. 人工神经网络的历史回顾

20 世纪 50 年代到 60 年代末期，人工神经网络发展迎来了第一次高潮，其重要成果是单级感知器及其电子线路模拟。单级感知器的研究成功，使人们过于乐观地认为找到了智能的关键，出现了对连接主义，尤其是对以感知器为代表的脑模型的研究热潮。

由于受到当时的理论模型、生物原型和技术条件的限制，脑模型研究在 20 世纪 70 年代后期至 80 年代初期落入低潮。直到 Hopfield 教授在 1982 年和 1984 年发表两篇重要论文，提出用硬件模拟神经网络以后，连接主义才又重新抬头。1986 年 Rumelhart 等人提出多层网络中的反向传播算法(BP)。此后，连接主义势头大振，从模型到算法，从理论分析到工程实现，为神经网络计算机走向市场打下了基础。

20 世纪 90 年代后，人们发现人工神经网络还有许多没有解决的问题，这其中包括许多理论问题，人工神经网络进入再认识和改进现有模型及算法的应用研究期。

二、人工神经网络基础

人工神经网络源于对生物神经网络的研究，所以以下介绍一下生物神经网络，另外还将介绍人工神经网络的基本知识：人工神经元模型、人工神经网络模型和人工神经网络的学习。

1. 生物神经网络

神经系统的基本构造是神经元(神经细胞)，它是处理人体内各部分之间相互信息传递的基本单元。每个神经元都由一个细胞体(或称为胞体)、一个轴突和一些树突(或称为枝蔓)组成。一个神经元的神经末梢(或称为轴突末梢)和其他神经元的细胞体或树突进行连接，连接点称为突触，突触是神经元之间的信息输入/输出接口，信息由一个神经元的神经末梢通过突触传递到另一个神经元的细胞体或树突(图 13-1)。

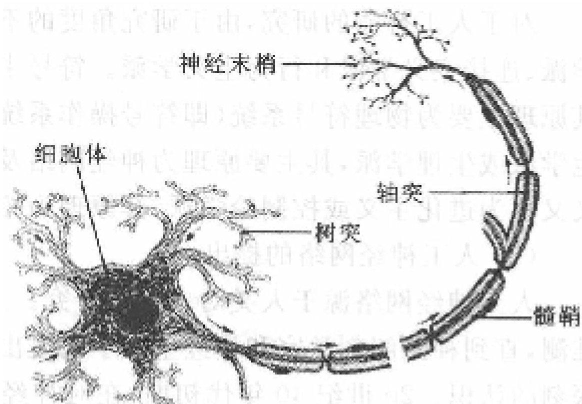


图 13-1 生物神经元示意图

人的大脑中约有 10^{11} 个生物神经元，它们通过 10^{15} 个连接组成一个生物神经网络系统。

每个生物神经元有独立接受、处理和传递电化学信号的能力。在突触的输入端，神经元细胞体对树突和细胞体各部位收到的来自其他神经元的信号进行接收和组合，这些

信号可以起到刺激或抑制的作用。当细胞体受到的刺激累加到一定程度后，细胞体就被激发，产生一个输出信号。输出信号沿轴突传到末梢，轴突末梢作为输出端通过突触将信号传给其他神经元。

在生物神经网络系统中，每个神经元通过突触与其他很多神经元相连接，接收来自其他神经元的不同信号，并向其他神经元发出信号，这些由同一个神经元发出的信号都是相同的。神经元接收到的信号强弱取决于它们之间突触的连接强度，连接强度越强，信号就越强，反之越弱。突触的连接强度是可以改变的。

2. 人工神经元结构

从模拟生物神经网络的角度出发，在人工神经网络中模拟生物神经元的是人工神经元(artificial neuron，AN)，其地位相当于生物神经网络的神经元。人工神经元是对生物神经元的抽象，它对生物神经元的结构和工作机制进行模拟。

(1) 人工神经元模型。

1943 年生理学家 McCulloch 和数理逻辑学家 Pitts 创立的阈值加权和模型，即 MP 模型是最早提出也是影响最大的人工神经元模型(图 13-2)。该模型基于对生物神经元工作机制的六点假设：

① 每个神经元都是一个多输入单输出的独立单元；

② 神经元接收到的信号有刺激和抑制两种类型之分；

③ 神经元具有空间整合特性和阈值特性；

④ 信号在神经元之间的传递有固定的时延；

⑤ 不考虑信号的整合作用时间和细胞体的反应时间；

⑥ 突触强度固定。

这里的阈值特性是指生物神经元受到的刺激累加到一定程度才能被激发。第六条假设是不符合实际情况的，这是因为受到当时认识的限制。

以上六点假设是对生物神经元工作机制的简化和抽象。通过这些假设，可以对生物神经元的工作机制进行形式化描述。

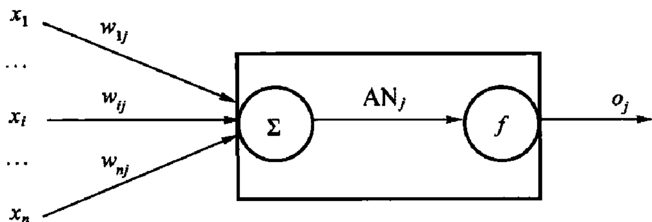


图 13-2 神经元模型示意图

图 13-2 中方框代表神经元 AN_j；x_1，…，x_n 为输入信号，x_i 代表神经元 AN_i 传递给神经元 AN_j 的信号；w_{ij} 代表神经元 AN_i 与神经元 AN_j 的连接强度，称为加权系数或权系数(权值)；o_j 表示神经元 AN_j 的输出信号。

神经元 AN_j 对其接收到的所有信号进行整合，即计算输入信号的加权和。然后神经元 AN_j 通过激活函数 f 对整合后的加权和结果进行变换，以确定所有输入信号的总效果是否超过了神经元 AN_j 的阈值，如果超过了阈值，该神经元就处于激发状态，否则处于抑制状

态。神经元状态的不同决定了输出信号 o_j。

在以下表述中，一般用大写粗体字母代表向量，用小写字母代表其中一个元素，比如神经元 AN_j 的输出用 o_j 表示，神经网络的输出用 $\boldsymbol{O}$ 表示。

如果令向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$，向量 $\boldsymbol{W}_j=(w_{1j},w_{2j},\cdots,w_{nj})^{\mathrm{T}}$，将神经元 AN_j 的输入信号的加权和表示为 net_j，那么

$$net_j = \boldsymbol{XW}_j \tag{13-1}$$

神经元 AN_j 的输出信号为：

$$o_j = f(net_j) = f(\boldsymbol{XW}_j) \tag{13-2}$$

(2) 激活函数。

激活函数也称为变换函数或激励函数、活化函数。激活函数通过控制神经元的输出，使神经元具有不同的信息处理特性，从而获得更广的适用范围。最为常见的激活函数有四种：阈值激活函数、线性激活函数、非线性激活函数、概率激活函数。

① 阈值激活函数。

阈值激活函数定义为：

$$f(net) = \begin{cases} \alpha & net > \theta \\ -\beta & net \leqslant \theta \end{cases} \tag{13-3}$$

式中 α,β,θ 均为非负常数，θ 为阈值。通常 α 取 1，β 取 0 或 -1。当 $\alpha=1,\beta=0,\theta=0$ 时称为单极性阈值激活函数；当 $\alpha=1,\beta=1,\theta=0$ 时称为双极性阈值激活函数。阈值激活函数如图 13-3 所示。

② 线性激活函数。

线性激活函数的一般形式为：

$$f(net) = k \cdot net + b \tag{13-4}$$

式中 k,b 均为常数。线性激活函数如图 13-4 所示。

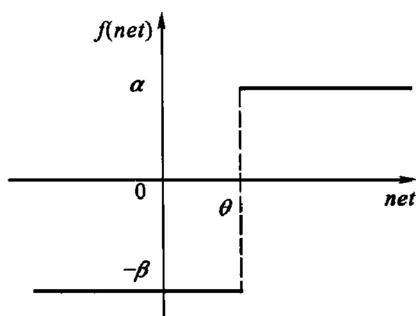


图 13-3　阈值激活函数

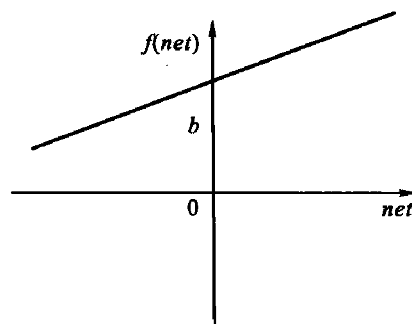


图 13-4　线性激活函数

③ 非线性激活函数。

非线性激活函数有 S 型激活函数和非线性斜面激活函数。

S 型激活函数是应用最广泛的一种激活函数，其一般形式为：

$$f(net) = a + \frac{b}{1+\exp(-d \cdot net)} \tag{13-5}$$

式中 a,b,d 均为常数。当 $a=0,b=1,d=0$ 时称为单极性 S 型激活函数；当 $a=-1,b=2$，

$d=0$时称为双极性 S 型激活函数。S 型激活函数如图 13-5 所示。

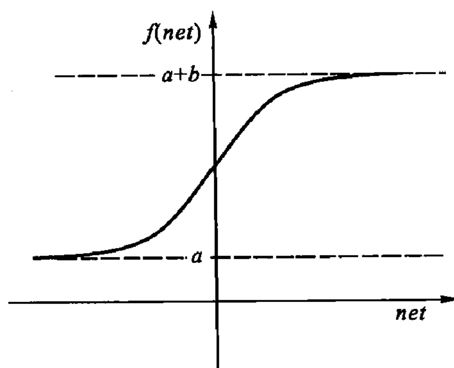


图 13-5 S 型激活函数

非线性斜面激活函数又称为分段线性激活函数或伪线性激活函数，其一般形式为：

$$f(net)=\begin{cases}\alpha & net \geqslant \theta \\ k \cdot net & -\theta < net < \theta \\ -\beta & net \leqslant -\theta\end{cases} \qquad (13\text{-}6)$$

式中 k,α,β 均为非负常数。当 $\alpha=1,\beta=0$ 时称为单极性分段线性激活函数；当 $\alpha=1,\beta=1$ 时称为双极性分段线性激活函数。非线性斜面激活函数如图 13-6 所示。

④ 概率激活函数。

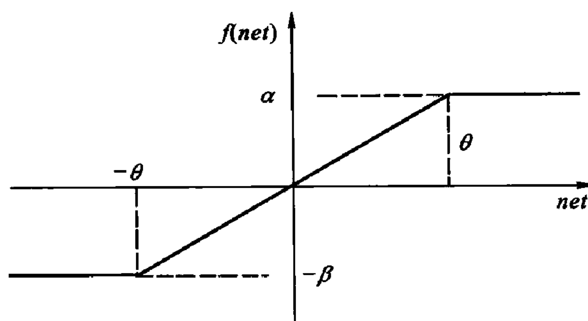


图 13-6 非线性斜面激活函数

概率激活函数的输出与输入没有确定的关系，其输出一般为 1 或 0，输出为 1 的概率为：

$$p(1)=\frac{1}{1+\exp(-net/T)} \qquad (13\text{-}7)$$

式中 T——温度参数。

3. 人工神经网络结构

(1) 人工神经网络的基本结构。

人工神经网络具有分层的结构。一般来讲，一个完整的人工神经网络由输入层、输出层和它们之间的隐藏层组成，隐藏层可以有多层，也可以没有。各层都可以含有数量不等的神经元，各层之间的连接方式也可以多种多样。

对于某个层内的一个神经元，它可以与同一层内的其他神经元连接，称为侧连接，也可与其他层的神经元连接，称为层内连接，甚至可以与它自己相连，这样就可以使每次的输出相关联，称为自连接或循环连接。当相互连接的两个神经元传递的信号起刺激作用时，把它们的连接强度表示为正实数，否则为负实数。

图 13-7 是两个神经元连接的模型，图中小圆圈表示一个神经元，箭头表示信息传递的方向，w_{ij} 表示神经元 AN_i 和 AN_j 的连接强度。

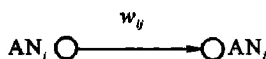


图 13-7 神经元连接模型

可以从信号传递的方向来考察神经网络中各层的关系。我们把信号在同一层内的侧向传递称为横向反馈，层间的向前传递称为层前反馈（简称前馈），层间的向后传递称为层反馈。横向反馈和层反馈统称为反馈，如图 13-8 所示。

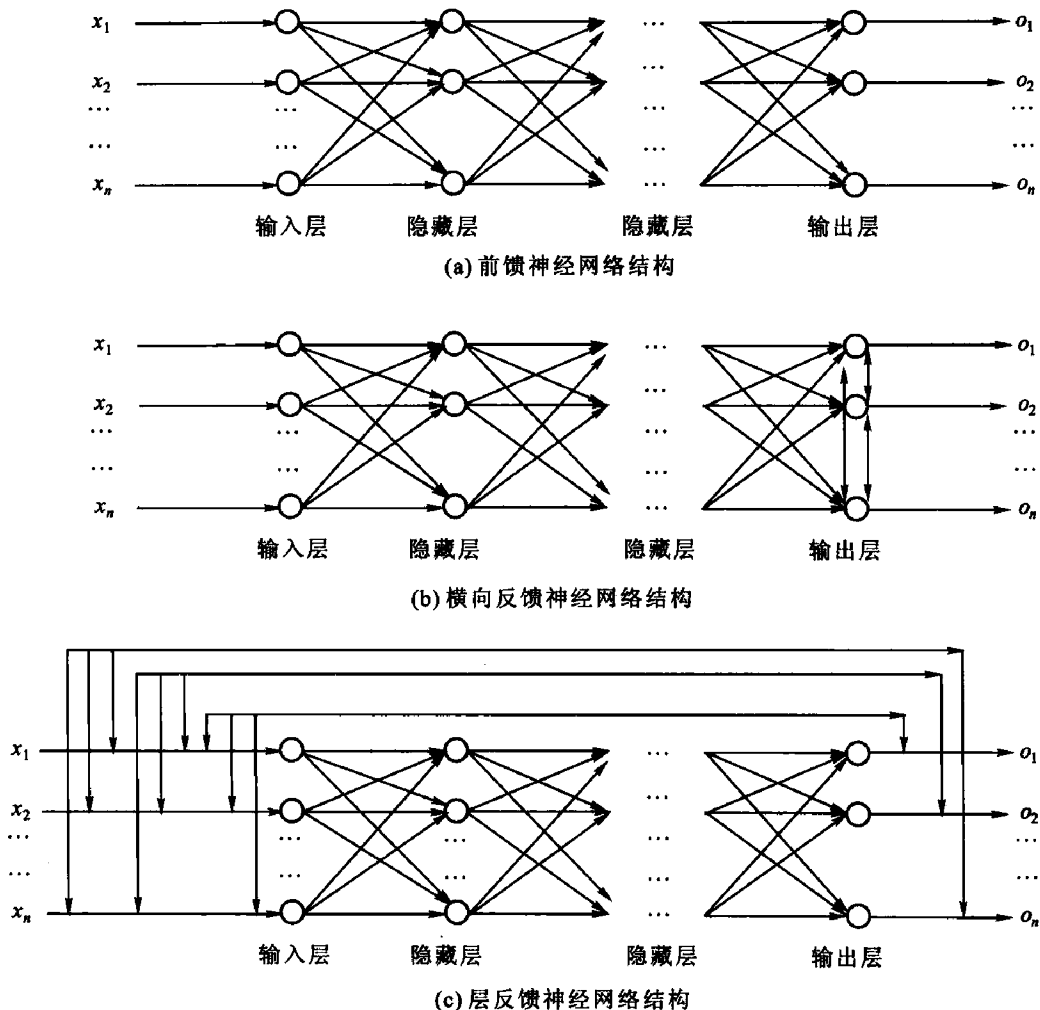


图 13-8 神经网络的基本结构

前馈神经网络结构中，输入层接收输入信号，输出层输出结果，其他相邻的两层，前一层的输出为下一层的输入。任何介于输出层和输入层之间的层都称为隐层，隐层的有无或多少都视需要而定。一般输入层只是起到将接收到的信息扇出的作用。任意相连的神经元之间都由连接强度表示它们之间的关系。

横向反馈神经网络结构中，侧向连接的层可以是任一层，也可以有多个侧向连接层，其他与前馈神经网络结构相同。

层反馈神经网络结构中，向后传递信号的层一般是输出层，接收反馈信号的层一般是输入层，这样的神经网络称为循环网络。循环网络中当前的输出会受到上次输出的影响，而上次输出受上次输入的影响，这样就形成一个迭代。在迭代的过程中，输入的原始信号被逐渐加强和修复。网络的输出可能每次都不相同，我们希望通过这种循环使得网络输出的信号趋于稳定。当网络的输出符合要求，一般指输出无变化时，我们称网络达到了平衡，循环也就可以停止了，否则称网络是不平衡的。如果网络始终无法平衡，那么称网络是无法收敛的。

实际的神经网络可以由以上三种基本结构组合而成，视具体需要而定。

(2) 激活函数。

人工神经网络不但需要结构的支撑，还需要激活函数来处理信息。

假定一个前馈神经网络有 n 层，$\boldsymbol{X}$ 是输入向量，$\boldsymbol{O}_i(1\leqslant i\leqslant n)$ 为第 i 层的输出向量，$\boldsymbol{W}^i(1\leqslant i\leqslant n)$ 为第 i 层到其后一层的连接矩阵，$\boldsymbol{NET}_i(1\leqslant i\leqslant n)$ 为第 i 层的网络输入向量，$F_i(1\leqslant i\leqslant n)$ 为第 i 层的激活函数，那么

$$\boldsymbol{O}_i = F_i(\boldsymbol{NET}_i) = F_i(\boldsymbol{XW}^i) = \boldsymbol{NET}_{i+1} \qquad (1\leqslant i\leqslant n) \tag{13-8}$$

网络最终的输出为：

$$\boldsymbol{O}_n = F_n(F_{n-1}(\cdots(F_2(F_1(\boldsymbol{XW}^1)\boldsymbol{W}^2))\cdots)\boldsymbol{W}^{n-1}) \tag{13-9}$$

神经网络每一层的激活函数都可以不同，但如果全部都是线性的激活函数，那么可以证明整个神经网络的功能只是相当于一个只有输入层和输出层的简单线性前馈神经网络。所以要实现复杂一些的功能，需要使用非线性激活函数。

4. 人工神经网络的学习

在 MP 模型中的第六条假设是“突触强度固定”，这与模拟人脑神经网络的目标相背，实际上生物神经元可以通过学习来改变它们之间的连接强度。

1949 年心理学家 Hebb 提出了著名的 Hebb 学习律。Hebb 认为如果源神经元和目的神经元均被激活，它们之间突触的连接强度就会增强，这被认为是最早最著名的 Hebb 学习律的生理学基础。

人工神经网络的学习是通过训练实现的，即将由样本向量构成的样本集合输入到人工神经网络中，然后按照一定的规则调整神经元之间的连接权，从而使神经网络的连接权矩阵存储样本集合的内涵，这就是神经网络学习的本质。

(1) 学习的形式。

从形式上可以将神经网络的学习分为有导师学习和无导师学习。

① 有导师学习。

有导师学习又称为有监督学习，它采用纠错的方式实现神经网络的学习过程。有导师学习不但要求给出输入向量，还需要给出相应的理想输出，每个训练样本都由这两部分组成。神经网络将其实际输出与理想输出相比较，如果不符合要求，神经网络就会按一定方式调整连接权值，使得神经网络的输出向理想输出靠近。重复这个过程，直到对样本集来说神经网络的输出达到一定的要求为止。

有导师学习算法常采用 Delta 规则。Delta 规则可以表示为：

$$w_{ij}(t+1) = w_{ij}(t) + \alpha[y_i - a_j(t)]o_i(t) \tag{13-10}$$

式中 $w_{ij}(t+1), w_{ij}(t)$——在时刻 $t+1$ 和 t 或第 $t+1$ 次和第 t 次训练时，神经元 AN_i 到神经元 AN_j 的连接强度；

$o_i(t)$——神经元 AN_i 在时刻 t 的输出；

α——给定的学习率；

y_i——神经元 AN_i 的理想输出；

$a_j(t)$——t 时刻神经元 AN_j 的激活状态，这里我们不区分激活状态和实际输出，可以认为是实际输出 o_j；

$y_i - a_j(t)$——导师因子或学习信号。

这里的学习率是给定的，代表了神经网络接受新信息的能力，其值应该设置在合理的范围内，如果太大，网络很难收敛，不容易趋于稳定；如果太小，学习的效率不明显。对于网络输入层的神经元，它们接收的信号不是来自其他神经元，而是输入信号，这时式(13-10)中的 $o_i(t)$ 应为 $x_i(t)$。

常见的有导师学习算法有感知器学习算法、BP 算法、最小均方学习算法、相关学习算法等。

② 无导师学习。

在无导师学习中，用来训练的样本只有输入向量。网络通过训练修改连接权矩阵，以使得网络能根据网络的内部结构和学习规则，在输入信息中发现可能存在的模式和规律。网络根据其功能和输入信息调整权值，这个过程称为网络的自组织。

Hebb 学习算法是一种著名的无导师学习算法，这种算法可以表示为：

$$w_{ij}(t+1) = w_{ij}(t) + \alpha o_j(t) o_i(t) \tag{13-11}$$

常见的无导师学习算法有 Hebb 学习算法、竞争与协同学习算法、随机连接算法、胜者为王算法等。

(2) 常见学习算法举例。

1990 年日本著名神经网络学者 Amari 提出了一种神经网络权值调整的通用学习规则，如图 13-9 所示。

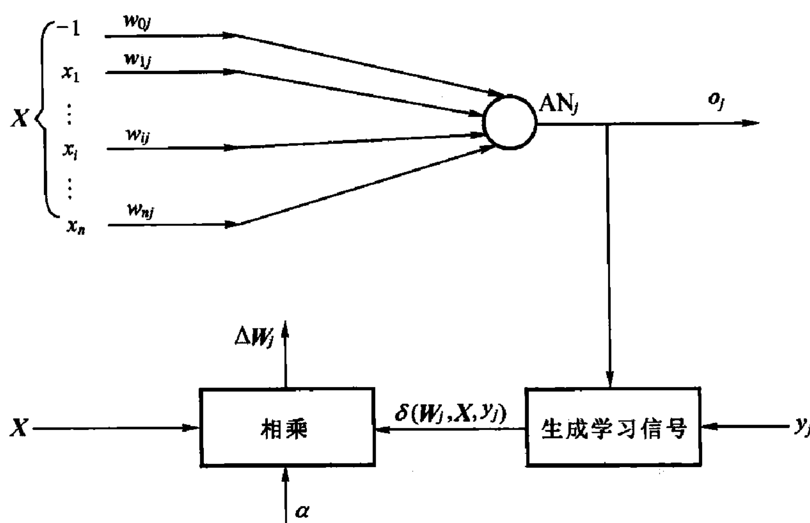


图 13-9　权值调整一般规则示意图

图中的神经元 AN_j 是网络中的某个节点；$\boldsymbol{X}$ 为该节点输入向量，它可以是网络的输入，也可以是其他节点的输出；w_{ij} 为第 i 个输入与神经元 AN_j 的连接权值，所有的连接权值构成了连接矩阵 $\boldsymbol{W}_j$，神经元 AN_j 的阈值被设定为 w_{oj}，且输入分量 x_o 恒为 -1；$\delta(\boldsymbol{W}_j, \boldsymbol{X}, y_j)$ 为生成的学习信号，或称为导师因子，它是实际输出 o_j 与理想输出 y_j 的函数，而 o_j 是 $\boldsymbol{W}_j$ 与 $\boldsymbol{X}$ 的函数，所以 δ 为 $\boldsymbol{W}_j, \boldsymbol{X}, y_j$ 的函数。连接矩阵的调整量 $\Delta\boldsymbol{W}_j$ 与 $\delta(\boldsymbol{W}_j, \boldsymbol{X}, y_j)$、学习率 α、输入向量 $\boldsymbol{X}$ 成正比。用数学公式表示为：

$$\boldsymbol{W}_j(t+1) = \boldsymbol{W}_j(t) + \alpha\delta(\boldsymbol{W}_j(t), \boldsymbol{X}(t), y_j)\boldsymbol{X}(t) \tag{13-12}$$

$$\Delta\boldsymbol{W}_j = \alpha\delta(\boldsymbol{W}_j(t), \boldsymbol{X}(t), y_j)\boldsymbol{X}(t) \tag{13-13}$$

式中的 $X(t)$ 即可以是来自网络外部的输入，也可以是其他节点的输出，所以式(13-10)和(13-11)都是式(13-12)的特例。

对于无导师学习，公式中不存在理想输出项 y_i。

学习信号 $\delta(W_j, X, y_i)$ 的定义不同，形成了各种各样的神经网络学习规则，下面简要介绍其中比较基本和重要的，对于其他更多更详细的内容请参考其他资料。

① Hebb 学习律。

前面已经多次提到过 Hebb 学习律，其主要内容为当连接的两个神经处于相同状态时，它们之间的连接强度应当增强，如果处于不同的状态，连接强度应当减弱。

在 Hebb 学习律中，学习信号等于神经元的输出，即

$$\delta = f(XW_j) \tag{13-14}$$

连接矩阵的调整向量为：

$$\Delta W_j = \alpha f(XW_j)X \tag{13-15}$$

这种 Hebb 学习律代表了纯前馈、无导师学习。Hebb 学习律要求权值有初始值，一般取零附近的随机小数。为防止输入与输出符号始终一致而出现权值无约束增长，应设定权值的饱和值。

【例 1】 设有四输入单输出神经网络，其阈值为 0，学习率 $\alpha=0.5$。样本向量 $X^1=(1,-1,1,-1)$，$X^2=(1,0,1,0)$，$X^3=(1,-1,0,0)$，初始连接矩阵为 $W(0)=(0.1,0,-0.2,0.3)^T$，使用双极性阈值激活函数求取 $W(3)$。

解：所谓双极性阈值激活函数，是当 net 大于阈值时，$f(net)=1$；当 net 小于等于阈值时，$f(net)=-1$（请参考阈值激活函数的介绍）。

输入第一个样本 X^1，有：

$$net^1 = X^1W(0) = (1,-1,1,-1)(0.1,0,-0.2,0.3)^T = -0.4$$

$$f(net^1) = -1$$

$$\begin{aligned} W(1)^T &= W(0)^T + \alpha f(net^1)X^1 \\ &= (0.1,0,-0.2,0.3) + 0.5\times(-1)\times(1,-1,1,-1) \\ &= (-0.4,0.5,-0.7,0.8) \end{aligned}$$

输入第二个样本 X^2，有：

$$net^2 = X^2W(1) = (1,0,1,0)(-0.4,0.5,-0.7,0.8)^T = -1.1$$

$$f(net^2) = -1$$

$$\begin{aligned} W(2)^T &= W(1)^T + \alpha f(net^2)X^2 \\ &= (-0.4,0.5,-0.7,0.8) + 0.5\times(-1)\times(1,0,1,0) \\ &= (-0.9,0.5,-1.2,0.8) \end{aligned}$$

输入第三个样本 X^3，有：

$$net^3 = X^3W(2) = (1,-1,0,0)(-0.9,0.5,-1.2,0.8)^T = -1.4$$

$$f(net^3) = -1$$

$$\begin{aligned} W(3)^T &= W(2)^T + \alpha f(net^3)X^3 \\ &= (-0.9,0.5,-1.2,0.8) + 0.5\times(-1)\times(1,-1,0,0) \\ &= (-1.4,1.0,-1.2,0.8) \end{aligned}$$

② 离散感知器学习规则。

感知器的学习是有导师学习。1958 年美国学者 Frank Rosenblatt 提出了一个具有单层计算单元的神经网络结构，称为感知器。实际上，单输出的感知器就是一个人工神经元(图 13-2)，而多输出的感知器类似于图 13-8 (a)，只不过没有中间的隐藏层，因为第一层输入层只是起到扇出输入信号的作用，所以仍可认为是一种单层的神经网络。

感知器的学习规则是学习信号等于神经元理想输出与实际输出之差，即：

$$\delta = y_i - o_i \tag{13-16}$$

感知器采用双极性阈值激活函数，连接矩阵的调整向量为：

$$\Delta \boldsymbol{W}_j = \alpha[y_j - f(\boldsymbol{X}\boldsymbol{W}_j)]\boldsymbol{X} \tag{13-17}$$

式中 y_j 和 $f(\boldsymbol{X}\boldsymbol{W}_j)$的取值为 1 或－1。

这种采用阈值函数作为激活函数的网络称为二值网络，其初始权值可取任意值。

③ 连续感知器学习规则。

如果把感知器的激活函数换成是连续可导的函数，比如 S 型激活函数，那么就称为连续感知器。1986 年心理学家 McClelland 和 Rumelhart 在神经网络的训练中引入了 ε 规则，该规则规定：

$$\delta = (y_i - o_i)f(\boldsymbol{X}\boldsymbol{W}_j) = [y_i - f(\boldsymbol{X}\boldsymbol{W}_j)]f'(\boldsymbol{X}\boldsymbol{W}_j) \tag{13-18}$$

这时的学习信号称为 ε。ε 规则保证了输出值与理想值的最小平方误差。由于连续感知器采用连续函数作为激活函数，所以 ε 规则可以推广到多层前馈网络，初始权值可取任意值。

三、BP 神经网络

BP 神经网络是目前应用最为广泛的一类神经网络，我们将介绍最基本的 BP 算法。

在上面提到的感知器算法中，理想输出与实际输出之差被用来估计连接权值的误差，但并没有考虑多级神经网络中隐藏层权值的调整，因为隐藏层的神经元没有理想的输出，难以估计其误差。

BP 神经网络是指采用 BP 算法的神经网络，为一非循环多级网络。BP 算法的含义是误差反向传播(error back-propagation)，正如其名字显示的那样，它利用输出层的误差来估计输出层的直接前导层的误差，再用这个误差估计更前一层的误差，以此类推，直到得到所有层的误差。

1. 网络的构成

BP 神经网络是非循环多级网络，可以有多个隐层。由于增加隐层层数和隐层神经元个数并不一定能够提高神经网络的精度和功能，而且会使网络变得复杂，所以一般情况下 BP 神经网络的隐层只有一层。

BP 算法要求所使用的激活函数必须是处处可导的，一般使用 S 型激活函数。假设一个神经元的输入为：

$$net = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n$$

那么其输出为：

$$o = f(net) = \frac{1}{1 + \mathrm{e}^{-net}}$$

o 对于 net 的导数为：

$$f'(net) = \frac{\mathrm{e}^{-net}}{(1 + \mathrm{e}^{-net})^2} = o(1 - o)$$

这里需要注意的是对于神经元的输出，当 *net* 落在(−1,1)之间时，o 变化较快；当 *net* 落在(−1,1)之外时，o 变化较慢(图 13-10)。所以当网络训练时需要对数据进行预处理，否则 *net* 可能很大，o 始终为 1，导致网络训练失败。

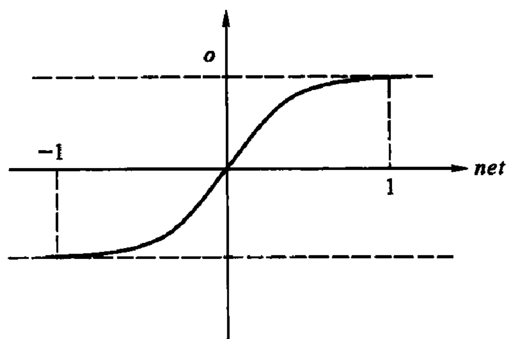


图 13-10 BP 网络神经元输出

2. 网络的训练

BP 网络神经的训练是有导师训练，其样本集由输入向量和理想输出向量构成。为保证网络训练的成功，初始的 BP 网络神经连接权值应该是一些零附近的随机小数。

BP 算法分为四个步骤：

① 将样本集中的一个样本($\boldsymbol{X}_i$,$\boldsymbol{Y}_i$)输入网络。

② 按式(13-9)计算网络输出 $\boldsymbol{O}_i$。

③ 得到理想输出 $\boldsymbol{Y}_i$ 与实际输出 $\boldsymbol{O}_i$ 之差。

④ 按极小化误差的方式调整权矩阵。

这个过程将一直循环下去，直到整个样本集的误差满足要求。这里将样本($\boldsymbol{X}_i$,$\boldsymbol{Y}_i$)的误差定义为：

$$E_i = \frac{\sum_{j=1}^{m}(y_{ij} - o_{ij})^2}{2} \tag{13-19}$$

式中 m——理想输出向量中元素的个数，即理想输出的个数，也就是说对于一个输入向量，它对应的输出也是一个向量，这时网络输出层的神经元个数也应该为 m。

整个样本集的误差为：

$$E = \sum_{i=1}^{n} E_i \tag{13-20}$$

3. 误差反向传播和权值调整

假设图 13-11 所示的网络为 BP 网络的一部分，第 $k-2$ 层和 $k-1$ 层的其他神经元没有表示出来，第 k 层有 m 个神经元。图中神经元 AN_i 与 AN_j 以 v_{ij} 连接，神经元 AN_j 到第 k 层的连接向量为(w_{j1},w_{j2},…,w_{jm})。δ_{jk-1} 表示神经元 AN_j 的输出误差，δ_{jk-1} 中的 $k-1$ 表示的是第 $k-1$ 层，j 表示的是第 j 个神经元，同样 δ_{1k} 表示第 k 层的第 1 个神经元的输出误差。

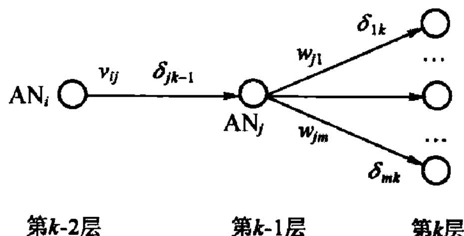


图 13-11 误差反向传播示意图

如果第 k 层为输出层，根据式(13-12)和(13-13)，有：

$$\Delta w_{jp} = \alpha\delta_{pk}o_j \tag{13-21}$$

式中　p——第 k 层中的某个神经元节点。

为保证了输出值与理想值的最小平方误差，根据 ε 规则，由式(13-18)得：

$$\delta_{pk} = f'_k(net_p)(y_p - o_p) \tag{13-22}$$

由于 $f'(net)=o(1-o)$，有：

$$\begin{aligned}\Delta w_{jp} &= \alpha\delta_{pk}o_j \\ &= \alpha f'_k(net_p)(y_p - o_p)o_j \\ &= \alpha o_p(1 - o_p)(y_p - o_p)o_j\end{aligned} \tag{13-23}$$

如果第 k 层不是输出层，可以假定其为输出层的直接前导层，以此类推，可以假定 w_{j1}，w_{j2}，…，w_{jm} 已经调整，δ_{1k}，δ_{2k}…，δ_{mk} 已知，如果可以表示 $\delta_{1,k-1}$，那么就可以调整整个网络的任何连接矩阵。BP 算法认为，δ_{1k}，δ_{2k}…，δ_{mk} 中的任何一个都有 $\delta_{1,k-1}$ 的作用，$\delta_{1,k-1}$ 是与 δ_{1k}，δ_{2k}…，δ_{mk} 有关的，即 $\delta_{1,k-1}$ 通过权 w_{j1} 作用于 δ_{1k}，通过权 w_{j2} 作用于 δ_{2k}，以此类推，通过权 w_{jm} 作用于 δ_{mk}。因此，可以用 $w_{j1}\delta_{1k}+w_{j2}\delta_{2k}+\cdots+w_{jm}\delta_{mk}$ 来表示 AN_i 的理想输出与实际输出之间的差，根据式(13-22)，有：

$$\delta_{j,k-1} = f_{k-1}'(net_j)(w_{j1}\delta_{1k} + w_{j2}\delta_{2k} + \cdots + w_{jm}\delta_{mk}) \tag{13-24}$$

又由于 $f'(net)=o(1-o)$，有：

$$\begin{aligned}\Delta v_{ij} &= \alpha\delta_{j,k-1}o_i \\ &= \alpha o_j(1 - o_j)(w_{j1}\delta_{1k} + w_{j2}\delta_{2k} + \cdots + w_{jm}\delta_{mk})o_i\end{aligned} \tag{13-25}$$

虽然 BP 网络取得了巨大的成功，但是它的一些问题也是非常明显的，比如训练速度慢、局部极小值问题、算法收敛问题等。对于这些问题的解决不再叙述，请参考其他资料。

四、神经网络在石油勘探及开发中的应用简介

由于神经网络具有其他方法无法比拟的优点，它在石油勘探及开发中取得了广泛的应用，主要有以下几点：

1. 预测渗透率

渗透率是石油地质勘探及开发的关键参数。实际上对渗透率的精确预测是很困难的。传统的方法主要是进行回归分析，通过建立孔隙度和渗透率的相关关系式，用孔隙度的资料来预测渗透率。用这种方法预测的结果最终忽视了最大值和最小值。与此相反，神经网络系统可以预测渗透率的精确变化，所用资料既可来自钻井岩芯也可来自测井资料。

BP 网络对预测渗透率较为有效。首先使用孔隙度值作为输入层，渗透率值作为输出层。这里须指出的是，输出层要包括样品的位置及计算点上下相邻的几十个孔隙度值，最终仅输出一个渗透率值。然后再移动所要计算渗透率的点位，同样输出坐标及该点上下相邻的几十个孔隙度值，然后再输出一个渗透率值。如此反复，便可得出孔隙度与渗透率的非线性对应关系。

2. 应用自组织特征映射网络识别储层

目前石油勘探在确定储层及测试层位时主要依据测井解释。由于测井响应与储层之间的关系十分复杂，很难用一种方程表达，而决定储层产油气与否的影响因素就更多，判断就更加困难。究其原因是测井解释模型仍未摆脱均质地层和线性方程这一致命的不可靠假设，以致解释结果不能令人满意。而自组织特征映射神经网络能完成输入与输出之间复杂的非线性映射，通过对简单的非线性函数进行迭合，便可实现复杂函数关系的转换，加上神

经网络所具有的完善的学习功能，自组织、自适应及联想记忆能力，以独特的信息处理方式为利用测井资料预测和判断储层的智能化提供了一个新的方法。

采用测井曲线上反映储层的特征参数训练网络的目的，在于对未知样本的准确预测，因为学习样本的正确性和准确性直接影响着预测模型的可靠性及精度。针对测井处理工作特点，特选定与储层物性、油气等特征密切相关，且在测井曲线上容易获取的电阻率、孔隙度、渗透率、泥质含量、声波、束缚水饱和度等参数，并依据数值大小分级，按照它们反映含油气储层的贡献强弱赋值，将参数进行归一化处理。

已知样本的特征参数经归一化赋值处理后，输入自组织特征映射网络进行训练学习和建模。输入的参数没有任何规律，但经网络训练后输出的各特征参数所代表的油层、油水层、水层、干层在平面上的分布呈局部集中，经划分区间后界限明显。自组织特征映射网络能利用从测井曲线上提取的特征参数将储集层的类别划分开。该网络模型还可以利用样本的测试结果对未测试层段的产量范围进行预测。

3. 自动识别岩性

利用反向传播算法可识别岩性。这种方法对测井解释岩性较为有效。输入层为声波时差自然电位、电阻率及自然伽玛曲线等测井曲线的特征值，输出层为砂岩、泥岩及灰岩等的期望值。在具体计算过程中，输入层及隐藏层的多少通常凭经验获得，并没有严格的规则可循。

神经网络经训练后，便将已知深度的测井曲线赋予相应的输入神经元，这些值通过网络到达输出层，之后输出层就能识别出测井曲线上的输入值代表的特定岩性。通过选取岩类及输出神经元，便可识别岩性。

这种方法优于传统的图形交会法和统计法，具有良好的识别能力，它不需要像统计法那样复杂精细的"预处理"，并且有较高的容错性、方便性及良好的适应性。

4. 描述油气藏非均质性

油气藏非均质性的主要参数有含油气岩石的孔隙度、渗透率、油气水饱和度。获得这些参数的主要途径有实验室测定、测井解释及统计方法。神经网络的出现为这些重要参数的预测提供了更为可靠的途径。在实际应用中可将深度、伽玛射线、体积密度及深感应测线作为神经网络的输入。通过网络的训练，可得出渗透率、孔隙度等参数的预测值。该方法能够为油藏模拟提供一个智能前端，它为地球物理测井的进一步利用提供了新的思维。

5. 进行地层对比

地层对比对研究岩性、岩相及油气横向连贯等研究有重要的意义。神经网络结合有序元素最佳匹配进行地层对比，可以克服各测井参数值的不规则给地层对比带来的不良影响，并可简化对比方法，降低工作量，从而达到提高地层对比的精确性。

这里须指出的是，在地层对比过程中，可将神经网络同经验及数学地质的其他方法相结合，进行综合对比分析。这些方法包括因子分析、马尔科夫链、最优分割法、聚类分析等。

6. 含油气有利性综合评价

利用人工神经网络进行石油勘探综合评价的基本原理是，通过若干已知井的含油气特征属性值（学习样本）对网络进行学习训练，使其获得评价专家的经验、知识，以及对评价指标的倾向性认识。当需要对未知井进行综合评价时，网络将再现专家的经验、知识和直觉思维，实现定性与定量的有效结合，保证评价的客观性和一致性。

7. 在油田开发中的应用

石油勘探及开发的最终目的是最大限度地提高石油采收率。运用人工神经网络"反向传播模型"的改进形式，将影响采收率的主要因素作为系统的输入，通过自学习算法，研究影响采收率的诸因素，最终建立油田采收率模型，能较好地描述各因素间的复杂非线性关系。

8. 在测井解释中的应用

测井曲线解释主要是为了解决以下两个问题：一是储层参数的选择和定量计算，二是识别问题，如油(气)水识别，岩性识别和裂缝识别等。

长期以来，传统的解释方法都建立在线性和均质测井理论基础上，将复杂问题进行简化假设，因而建立的测井解释模式常常与实际不符。实际上，测井响应与地层特性之间的关系是十分复杂的，很难用一种方程表达出来。而神经网络对解决这类问题有独特的优越性，它能完成输入与输出之间复杂的非线性映射。通过对简单的非线性函数进行几次复合，便可实现复杂函数关系的转换，加之神经网络具有较完善的学习功能、自适应能力、联想记忆能力以及独特的信息处理方式等，因此在测井定量解释中，不需要事先建立任何测井响应方程或提供经验公式，也避免了解释过程中一些中间参数选择、解释模型和测井响应方程的建立以及求解所带来的人为误差，从而大大提高了测井解释的精度和速度，为测井曲线解释开辟了一条新途径，为地层评价、地层结构构造及沉积环境研究、寻找油气储层提供了可靠的基础数据。

9. 在地震勘探中的应用

由于神经网络具有很强的优于常规模式的识别能力，它能通过样本进行学习，具有在高噪音干扰下的适应性，并能大量减少工作量。因此，它已被广泛用于地震勘探的各个环节中，如初始波自动收取，地震道自动编辑，地震资料自动解释，反射层自动追踪，地震亮点或异常振幅的自动检测，地震剖面模式识别及油气预测专家系统等方面。

§2 应用实例

【例 1】采用 BP 神经网络对川东大天池构造石炭系的天然气富集程度进行横向预测。资料来自该构造带上 14 口井的测井资料，选用 7 个储集层集总参数作为衡量天然气富集程度的特征参数：有效厚度、平均孔隙度、平均含气饱和度、储层综合判别集总参数、天然气体积集总参数、基质渗透率集总参数、裂缝总孔隙度集总参数。选取天东 26 井等共 11 口井的参数作为学习样本，以计算的单井储量丰度作为理想输出。选取天东 1 井等共 3 口井作为工作目标。首先将全部数据进行最大值标准化，然后将学习样本输入网络进行学习。程序在进行了 18 700 次训练后达到了精度要求，即样本集总误差小于 0.001。将预测的结果进行还原便得到最终的预测值。

需要说明的是由于初始的权矩阵是随机的，隐藏层神经元个数也可以不同，所以训练成功需要的次数会有变化，预测的结果也会有细微不同。为得到尽量准确的结果，训练样本应尽可能准确，当然大前提是选择正确的特征参数。

隐藏层神经元的节点数一般取 2×(输入层节点数＋输出层节点数)－1，本例中为 15。

数据及计算结果见表 13-1。

表 13-1 大天池构造石炭系天然气富集横向预测

井 号	有效厚度/m	平均孔隙度/%	平均含气饱和度/%	储层综合判别值集总参数 ΣCV	天然气体积集总参数 ΣCV	基质渗透率集总参数 ΣCV	裂缝总孔隙度集总参数 $\Sigma\varphi f$	计算的单井储量丰度/($\times 10^8$t·km^{-2})	神经网络预测值
天东 26	39	6.23	83.03	395	20.66	29.7	7.78	6.31	—
天东 16	25	7.86	82.3	315.1	16.35	79.66	13.51	5.72	—
天东 19	31.6	6.59	83.3	550.9	28.8	30	9.2	5.368	—
天东 12	27.4	5.8	82	205.6	12.3	70.9	6.2	4.27	—
大天 2	26	6.2	84.4	276.3	13.78	18.8	4	4.82	—
天东 9	31.5	4.35	79.5	195.4	11.1	6.7	4	3.57	—
天东 7	20.8	5	85	209.7	9.89	8.36	2.41	3.78	—
门西 1	9.25	4.67	79.52	61	3.46	2.15	1.12	1.57	—
天东 4	9.3	4.03	78.4	60.5	3.4	0.55	2.1	0.96	—
天西 2	3.75	4	77.8	18.7	0.2	0.53	0.81	0.38	—
大天 1	0.25	3.85	67.2	0.24	0.06	0.001	0.02	0.02	—
天东 1	24.3	8.05	81.2	301.3	16.11	59.64	14.76	5.64	5.77
天东 11	23	5.67	80	193	10.55	11.51	3.45	3.68	3.80
天东 22	14	5.3	77.5	99.6	5.79	2.59	2.54	2.02	2.33

注："—"表示训练样本。

【例 2】在回归分析一章中曾经分析了 18 个盆地的大量生油门限时间与生油层温度和埋藏深度之间的关系，在这里用 BP 神经网络再进行一次分析。

首先对数据进行最大值标准化，以生油层温度和埋藏深度为输入信息，以前 13 个盆地的生油门限时间为理想输出，进行网络训练，以后 5 个盆地为工作目标。输入层神经元节点数设为 2，输出层节点数设为 1，隐藏层节点数设为 5。

经过训练发现网络难以收敛，分析原因可能是学习样本中蕴含的规律并不明显或有利群样本。经过分析发现网络样本集误差稳定在 0.04 左右，故将误差精度设为 0.04，两个学习率都设为 0.2。网络在进行了约 95 000 次训练后成功，数据和工作目标预测结果见表 13-2。

表 13-2 神经网络预测生油门限时间

序 号	盆地名称	生油层温度/℃	埋藏深度/m	生油门限时间/Ma	预测生油门限时间/Ma	误差/%
1	杜阿拉盆地	65	1 200	70	—	
2	落山矶盆地	115	2 440	12	—	
3	文吐拉盆地	127	2 740	12	—	
4	巴黎盆地	60	1 400	180	—	
5	阿启坦盆地(1)	90	3 300	112	—	
6	阿启坦盆地(2)	72	2 500	135	—	
7	卡马尔圭盆地	106	3 250	38	—	

续表

序　号	盆地名称	生油层温度 /℃	埋藏深度 /m	生油门限时间 /Ma	预测生油门限时间 /Ma	误差 /%
8	阿尤恩盆地	85	2 740	105	—	
9	苏禄海盆地	120	3 050	12	—	
10	塔拉纳基盆地(1)	80	2 900	70	—	
11	亚马逊盆地	62	1 750	359	—	
12	塔拉纳基盆地(2)	95	3 350	32	—	
13	东营凹陷	93	2 200	35	—	
14	潜江盆地	90	2 200	35	43.798	25
15	松辽盆地(1)	70	1 330	110	54.568	50
16	松辽盆地(2)	65	1 230	100	83.647	16
17	松辽盆地(3)	63	1 180	90	97.289	8
18	辽河盆地	81	1 700	50	39.131	22
平均误差						24

注:“—”表示训练样本

与回归分析相比,可以发现神经网络计算结果的明显不同,这是由它们各自的原理所决定的。

参考文献

【1】李汉林，赵永军．石油数学地质．东营：石油大学出版社，1998

【2】康永尚，沈金松，谌卓恒．现代数学地质．北京：石油工业出版社，2005

【3】刘承祚．数学地质的主要进展和发展趋势．地质论评，1996，42(4)

【4】赵旭东．石油数学地质概论．北京：石油工业出版社，1992

【5】李公时，谢国柱．数学地质教程．长沙：中南工业大学出版社，1989

【6】武守城．石油资源地质评价导论．北京：石油工业出版社，1986

【7】刘承祚，孙惠文．数学地质基本方法及应用．北京：地质出版社，1982

【8】王学仁．地质数据的多变量统计分析．北京：科学出版社，1982

【9】Harbaugh J W．地质过程的计算机模拟．罗人彦，译．北京：地质出版社，1986

【10】孙洪泉．地质统计学及其应用．北京：中国矿业大学出版社，1990

【11】张俊福，邓本让，朱玉仙，等．应用模糊数学．北京：地质出版社，1988

【12】张跃，邹寿平，宿芬．模糊数学方法及其应用．北京：煤炭工业出版社，1992

【13】谢季坚，刘承平．模糊数学方法及其应用．武昌：华中科技大学出版社，2000

【14】石广仁．地学中的计算机应用新技术．北京：石油工业出版社，1999

【15】侯景儒，尹镇南，李维明，等．实用地质统计学．北京：地质出版社，1998

【16】王仁铎，胡光道．线性地质统计学．北京：地质出版社，1988

【17】《油气资源评价方法研究与应用》编委会．油气资源评价方法研究与应用．北京：石油工业出版社，1988

【18】于志钧，赵旭东．石油数学地质．北京：石油工业出版社，1986

【19】余金声，李裕伟．地质因子分析．北京：地质出版社，1985

【20】陆明德，天时芸．石油天然气数学地质．北京：中国地质大学出版社，1991

【21】方开泰，潘思沛．聚类分析．北京：地质出版社，1982

【22】李汉林，连承波，马士坤，等．基于气侧资料的储层含油气性识别方法．石油大学学报(自然科学版)，2006，30(4)：21-23

【23】李汉林，傅晓宁，翟汝霞．资源总量随机求和中存在的问题及改进方法．石油大学学报(自然科学版)，2007，31(6)：10-12

【24】李汉林，赵永军．用体积单元有机碳生烃律法预测油气资源量．大庆石油学院学报，1999，23(3)：5-7

【25】李汉林，赵永军．岩性识别的多元统计方法．地质论评，1998，44(1)：106-111

【26】诸克军，冯光华，冯刚顶．一种石油勘探决策的模糊聚类分析法．江汉石油学院学报，1999，21(3)：15-16

【27】黄竞先，侯景儒．泛克立格法和指示克立格法在地球化学探矿中的应用．地球科学-中国地质大学学报，1994，19(3)：321-327

【28】肖斌，潘懋，赵鹏大，等．时空多元指示克立格法的理论研究．北京大学学报(自然

科学版),2001,37(1):36-40

【29】连承波,赵永军,李汉林,等.煤层含气量的主控因素及定量预测.煤炭学报,2005,30(6):726-729

【30】赵永军,李汉林.油气地表化探中基于R型因子分析的综合指标及工区N级划分.物探化探计算技术,2000,22(3):233-237

【31】赵文智.石油地质理论与方法进展.北京:石油工业出版社,2006

【32】方朝亮.勘探开发集成配套技术及应用实践.北京:石油工业出版社,2006

【33】韩力群.人工神经网络教程.北京:北京邮电大学出版社,2006

【34】蒋宗礼.人工神经网络导论.北京:高等教育出版社,2006

【35】杨行峻,郑君里.人工神经网络.北京:高等教育出版社,1992

【36】杨铭震,王燕霞.人工神经网络及其在石油勘探中的应用.北京:兵器工业出版社,1993

【37】金燕.人工神经网络在测井地质领域中的应用.天然气勘探与开发,1999,22(1)

【38】陶淑娴,肖慈询,杨斌,等.神经网络在测井解释中的应用.石油勘探,1995,34(3)

【39】张玉池,温佩琳等.人工神经网络在地球物理勘探中的应用概论.矿产与地质,1999,13(6)

【40】曹思远,梁春生.储层预测中BP神经网络的应用.地球物理学进展,2002,17(1)

【41】傅强,王家林,周祖翼.自组织特征映射网络在储层识别中的应用.同济大学学报.1996,27(3)

【42】于建华.应用人工神经网络自动识别岩性.石油地球物理勘探,1993,28 (1)

【43】王向公,张超谟,黄文新,等.神经网络在地层对比中的应用.国外油气科技.1995,3(44)

【44】诸克军,杨久西等.基于人工神经网络的石油勘探有利性综合评价.系统工程理论与实践,2004,22(4)